Peer Community Journal

Section: Genomics

RESEARCH ARTICLE

Published
2023-10-13

Cite as
Grégoire Aubert, Jonathan Kreplak, Magalie Leveugle, Hervé Duborjal, Anthony Klein, Karen Boucherot, Emilie Vieille, Marianne Chabert-Martinello, Corinne Cruaud, Virginie Bourion, Isabelle Lejeune-Hénaut, Marie-Laure Pilet-Nayel, Yanis Bouchenak-Khelladi, Nicolas Francillonne, Nadim Tayeh, Jean-Philippe Pichon, Nathalie Rivière and Judith Burstin (2023) *SNP discovery by exome capture and resequencing in a pea genetic resource collection*, Peer Community Journal, 3: e100.

Correspondence
Gregoire.Aubert@inrae.fr

Peer-review
Peer reviewed and recommended by PCI Genomics, https://doi.org/10.24072/pci.genomics.100237

SNP discovery by exome capture and resequencing in a pea genetic resource collection

Grégoire Aubert[1], Jonathan Kreplak[1], Magalie Leveugle[2,3], Hervé Duborjal[2,3], Anthony Klein[1], Karen Boucherot[1], Emilie Vieille[1], Marianne Chabert-Martinello[1], Corinne Cruaud[4], Virginie Bourion[1], Isabelle Lejeune-Hénaut[5], Marie-Laure Pilet-Nayel[6], Yanis Bouchenak-Khelladi[1], Nicolas Francillonne[7,8], Nadim Tayeh[1], Jean-Philippe Pichon[2,3], Nathalie Rivière[2,9], and Judith Burstin[1]

Volume **3** (2023), article e100

https://doi.org/10.24072/pcjournal.332

Abstract

Pea is a major pulse crop in temperate regions and a model plant in genetics. Large genetic marker resources are needed to assess the genetic diversity in the species genepool and to provide selection tools for breeders. In this study, we used second-generation sequencing to perform an exome-capture protocol using a diverse pea germplasm collection, and produced a resource of over 2 million Single Nucleotide Polymorphisms. This dataset was then used to characterize the genetic diversity present in the panel and compute phylogenetic and structure analyses. The development of this resource paves the way for Genome-wide association studies and the development of powerful genotyping tools.

[1]Agroécologie, INRAE, Institut Agro, Univ. Bourgogne, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France, [2]Biogemma, F-63720 Chappes, France, [3]Present address: Limagrain F-63720 Chappes, France, [4]Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, 2 Rue Gaston Crémieux, 91057 Evry, France, [5]BioEcoAgro, INRAE, Univ. Liège, Univ. Lille, Univ. Picardie Jules Verne, 2, Chaussée Brunehaut, F-80203, Estrées-Mons, France, [6]IGEPP, INRAE, Institut Agro, Univ Rennes, 35653, Le Rheu, France, [7]Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France, [8]Université Paris-Saclay, INRAE, BioinfOmics, Plant bioinformatics facility, 78026, Versailles, France, [9]Present address: HM-Clause 49800 Loire-Authion

Background & Summary

In addition to being the model plant used by Mendel (1866) to establish genetic laws, pea (*Pisum sativum* L.) is a major pulse crop cultivated in many temperate regions of the world. In order to face new challenges imposed particularly by global climate change and new regulations targeted at reducing chemical inputs, pea breeders have to take advantage of the genetic diversity present in the *Pisum* genepool to develop improved, resilient varieties. The aim of this study was to assess the genetic diversity of a pea germplasm collection and allow genome-wide association studies using this collection.

To be able to perform genome-wide association approaches with high resolution, genotyping with a large set of genetic markers such as Single Nucleotide Polymorphism (SNP) markers well-spread over the genome is required. Rapid advances in second-generation sequencing technologies and the development of bioinformatic tools have revolutionized the access to and the characterization of available genetic diversity. High-density, high-throughput genotyping has been possible for a large number of species, including those with large and complex genomes (Hill et al., 2019) such as pea which genome size is estimated to be 4.45 Gb (Doležel et al., 1998). In this study, which is part of the PeaMUST project (Burstin et al., 2021), we used a target capture technology based on pea transcriptome sequences to generate exome-enriched genomic libraries that were further subjected to Illumina sequencing in paired-end mode. This methodology was chosen because whole-genome resequencing is relatively expensive for species with large genomes and because capturing genetic variations in repeated non-coding regions is difficult to achieve or to interpret (Ku et al., 2012). Whole-exome sequencing represented an interesting alternative that focused on coding regions only (Ng et al., 2009, 2010). Mapping the obtained reads on the reference pea genome sequence enabled the discovery of an abundant set of SNPs. The development of this resource is a crucial cornerstone in research and breeding projects towards boosting the improvement of pea production and quality.

Methods

Plant material and DNA extraction

A set of 240 Pisum accessions was selected, including 220 accessions originated from a larger panel structured into 16 genetic groups (Siol et al., 2017) and 20 additional accessions chosen for their phenotypes. The 240-accession collection is referred to as Architecture and Multi-Stress (AMS) collection, since the accessions represent a broad diversity for root and shoot architecture and for biotic and abiotic stress responses. This collection contains cultivars, landraces, and wild types (including some *Pisum fulvum* and *Pisum sativum* subspecies accessions) with diverse geographical origins (Supplementary Table 1). Leaves were collected from 10 plants per accession, flash frozen in liquid nitrogen and stored at -80°C prior to DNA extraction. Tissues were then ground in liquid nitrogen using a pestle and a mortar. Genomic DNA extraction was performed using Nucleospin PlantII minikit (Macherey-Nagel, Hoerdt, France) following the manufacturer's instructions.

Probe design

As the pea genome sequence was not available at the time the probe design was made, two pea transcriptome datasets (Duarte et al., 2014; Alves-Carvalho et al., 2015) were used to build a reference set (refset) of expressed genes. After redundancy was removed, 67,161 unique contigs were kept, 20,972 of them being common to the two sequence datasets. The first exome capture design based on the refset was undertaken after predicting putative exon/intron junctions, masking repetitive sequences as well as excluding putative mitochondrial and chloroplastic sequences. A first probe design was performed by Roche™ (Madison, WI, USA) targeting 68.3 Mb. The analysis of the first capture results with the original probes demonstrated that a minority of 10-15% of the contigs retained the majority of the sequencing efforts, resulting in insufficient coverage for the remaining contigs. Sequencing data were used to identify the repetitive regions and a new probe design was performed after masking them to target a final capture space of 41.3Mb, representing 51,225 cDNA contigs.

Library Preparation and target enrichment

DNA samples were normalized before being fragmented with Adaptive Focused Acoustics® Technology (Covaris Inc., Massachusetts, USA). A 250-bp target size was obtained by using a Covaris E220 system, according to the manufacturer's instructions. DNA fragments underwent then a NGS library preparation procedure consisting in end repair and Illumina adaptor ligation using the KAPA HTP kit (Roche, Basel, Switzerland). Individual index sequences were added to each library for identifying reads and sorting them according to their initial origins. The Sequence Capture was performed using SeqCap EZ Developer kit (Bainbridge et al., 2010) according to the manufacturer's instructions (Roche™). The sequence capture reaction efficiency was evaluated by measuring, using quantitative PCR, a relative fold enrichment and loss of respectively targeted and non-targeted regions before and after the sequence capture reaction.

Targeted resequencing, sequence alignment and SNP calling

The captured samples were sequenced on HiSeq 2000 sequencing platform (Illumina, California, USA) with a Paired-End sequencing strategy of 2 reads of 100 bases. The sequenced reads were trimmed for adaptor sequences using cutadapt 1.8.3 (Martin, 2011). Low-quality nucleotides with quality value<20 were removed from both ends. The longest sequence without adapters and low-quality bases was kept. Sequences between the second unknown nucleotide (N) and the end of the read were also trimmed. Reads shorter than 30 nucleotides after trimming were discarded. These trimming steps were achieved using fastx_clean (http://www.genoscope.cns.fr/fastxtend), an in-house software based on the FASTX library (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The reads were then aligned on the targeted regions with Novoalign V3.09 (http://www.novocraft.com, Selangor, Malaysia). We took advantage of the v1 genome sequence of cv. Cameor published meanwhile (Kreplak et al., 2019) to perform SNP detection. Single nucleotide variants were detected on all the samples using samtools mpileup, followed by bcftools call and bcftools filter (Li et al., 2009) with a minimum genotype quality of 20 and a minimum coverage of 5 reads per sample. SNP variants for which more than 10 percent of the samples in the panel were heterozygous were then filtered out.

Phylogenetic analysis

A subset of 206,474 SNPs was selected from the Dataset by applying filters on missing data (<20%), Minor Allele Frequency (>1%) and linkage disequilibrium (LD pruning based PLINK --indep 50 5 2). This subset was used to build a maximum likelihood phylogenetic tree using IQtree (Minh et al., 2020), version 2.1.2, with model of substitution GTR+F+ASC. Alternatively, we also used a coalescent approach using 10,000 SNP non-overlapping windows as described by Wang et al. (2022). The trees were also constructed using IQtree with the same parameters and finally were summarized using ASTRAL v5.15.1 (Zhang et al., 2016). Tree visualizations were generated using R package ggtree (Yu et al., 2017).

Structure analysis

Structure within the collection was calculated using the Bayesian clustering program FastStructure (Raj et al., 2014) using a logistic prior for K ranging from 1 to 10. The script chooseK.py (part of the FastStructure distribution) was used to determine the best K that explained the structure in the collection based on model complexity. In addition, discriminant analysis of principal components (DAPC) was applied using DAPC function from adegenet package (version 2.1.5) (Jombart et al., 2010) in order to describe the genetic structure of the panel. Both FastStructure and DAPC groups were visualized on the phylogenetic tree using R package ggtreeExtra (Yu et al., 2017).

Results

Data Records

The collection of 240 pea accessions was genotyped using an original set of probes for exome capture. Sequence data are available on NCBI (Bioproject PRJEB56612) and the number of sequencing reads per accession is given in Supplementary Table 2. Using the pea Cameor genomic sequence v1 (Kreplak et al., 2019) as a reference, 2,285,342 SNPs were identified. The full set of variants has been recorded as a VCF file and deposited at URGI (https://doi.org/10.15454/3QRIPA; Kreplak et al., 2021). The variant statistics per accession are reported in Supplementary Table 3. In average, 183,170 homozygous variants (compared

to Cameor genome sequence) were detected per accession. As expected, the genotyping data from Cameor (DCG0251) used as an accession in the AMS panel were overall conform to the reference alleles from the reference genome sequence. Differences were only seen for 121 positions (0,005%). Among the detected SNPs, 647,220 were singletons (detected in only one accession). Excluding the reference Cameor (DCG0251), the number of singletons per accession ranged between 2 for VKL0176, a cultivated winter fodder pea to 113,086 for VSD0034, a *Pisum sativum* subsp. *abyssinicum* accession. The set of identified polymorphisms spans the seven pseudomolecules of the Cameor genome assembly at a frequency varying from 1 variant every 1831 bases for chromosome 2 to one every 1345 bases for chromosome 1 (Supplementary Table 4).

Variant effects

We used the snpEff program (Cingolani et al., 2012) in order to categorize the detected SNPs according to their predicted effects or their locations (Supplementary Table 4). The vast majority (76,71%) were labelled as “Modifier” (placed upstream or downstream of genes) and 14.32%, 8.76%, and 0.22% of the SNPs had a low (no change of the protein sequence), moderate (change of amin oacid), and high (major change in the protein) predicted impact on gene functions, respectively. In fact, out of the total SNPs detected in coding regions (23.29%), 57.52% were predicted to be silent, 41.95% were predicted to induce an amino acid change in the coded protein and 0.53% were predicted to have a disruptive nonsense effect (premature stop codon, splicing junction modification or start codon missing).

Our data provide insights into exome genetic variation and highlight mutations with functional effects. This polymorphism inventory is valuable to explain the phenotypic diversity in the *Pisum* species.

Phylogeny and structure analysis

Three different complementary approaches (Discriminant analysis of principal components, FastSTRUCTURE, Maximum Likelihood Phylogenetic trees using both standard substitution model and a coalescence-based approach) have been used to study the structure of the germplasm panel.

DAPC led to a clustering of the accessions into seven groups, as the most likely structure according to Bayesian information criterion. Clustering (Supplementary Table 1, Figure 1) tended to separate accessions according to crop evolution and cultivation types. Cluster 1 consisted in 14 accessions, mainly landraces or primitive germplasm. Cluster 2 comprised 87 accessions with an important proportion of spring cultivars including garden peas. Cluster 3 is composed of 27 accessions with a majority of spring-type lines coming from an Aphanomyces root rot breeding programme from Groupement des Sélectionneurs de Protéagineux (GSP, France) and Cluster 4 was a mix of spring-type and winter-type field peas with three primitive or landrace accessions. Cluster 5 grouped 5 wild-type accessions including *Pisum fulvum*, and Clusters 6 and 7 included winter-type field pea and fodder pea cultivars, respectively.

FastStructure, on the other hand, inferred that the panel was divided into five ancestral subpopulations (numbered A to E, Figure 1). Subpopulation E corresponded to the DAPC cluster of 5 wild accessions with very little admixture with the other clusters while subpopulation C corresponded to Cluster 1 (landraces/primitive germplasm). Subpopulation D corresponded to the fodder pea cluster 7. Clusters 3, 4 and 6 (field pea cultivars mainly) seemed to derive from the ancestral subpopulation A. Some admixture between subpopulations A and D was observed for the winter field pea accessions from Cluster 6, and between subpopulations A, B and D for some garden pea cultivars.

The Phylogenetic tree (Figure 1, produced using the 206,474 SNPs) confirmed the DAPC and FastStructure observations with only few placement differences. The summarized phylogenetic tree inferred with the coalescent approach (Supplementary Figure 1) corroborated the clade delimitations. However, the relationships between clades differed (Supplementary Figure 2) which could be attributed to admixture during cultivar selection leading to incomplete lineage sorting.

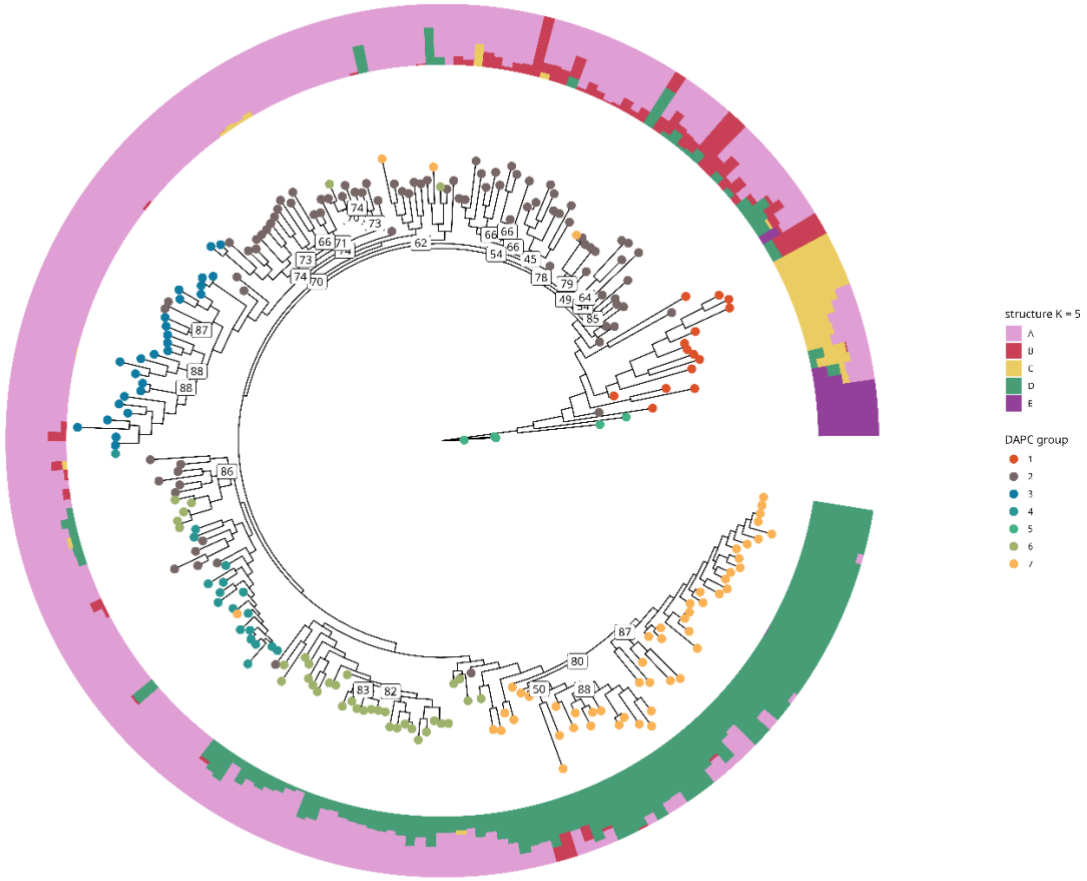

Figure 1 Maximum-likelihood phylogenetic analysis using 206,474 SNPs, DAPC grouping, and FastStructure composition of the 240-accession pea AMS collection. Colour codes indicate the different groups as inferred by DAPC and FastStructure. Numbers shown at the nodes show bootstrap support when below 90%.

Conclusion

In conclusion, this dataset is a large marker resource that can be used for different purposes, including the development of targeted genotyping tools for molecular identification, genetic mapping or genomic selection in pea. It provides insights into pea diversity and helps to investigate selection processes in this species. The SNP resource also empowers Genome-wide Association Studies targeted at revealing the genetic architecture of important traits and highlighting alleles to be used in pea breeding programmes. Indeed, the collection has been evaluated for different traits including plant architecture, phenology and resistance or tolerance to a range of biotic and abiotic stresses, as exemplified by Ollivier et al. (2022) who deciphered the genetic determinism of resistance to two aphid biotypes.

Acknowledgements

Preprint version 4 of this article has been peer-reviewed and recommended by Peer Community In Genomics (https://doi.org/10.24072/pci.genomics.100237; Nawae, 2023). The authors wish to thank the reviewers and the recommender for very useful comments on the manuscript.

Author Contributions

ILH, MLPN, VB and JB designed the diversity panel. Experiments were conceived and designed by GA, NR, JPP and JB, and performed by HD, AK, KB, EV, MCM and CC. Data were analysed by ML, JK, YBK and GA and NF organised the data access. The manuscript was written by GA, NT and JK and all authors read and improved it.

Funding

This study is part of the PeaMUST project, that was funded by the French government through the Investment for the Future program (project ANR-11-BTBR-0002) and Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data, scripts, code, and supplementary information availability

Data and Supplementary information are available online: https://doi.org/10.15454/3QRIPA (Kreplak et al., 2021).

References

Alves-Carvalho S, Aubert G, Carrère S, Cruaud C, Brochot A-L, Jacquin F, Klein A, Martin C, Boucherot K, Kreplak J, da Silva C, Moreau S, Gamas P, Wincker P, Gouzy J, Burstin J (2015) Full-length *de novo* assembly of RNA-seq data in pea (*Pisum sativum* L.) provides a gene expression atlas and gives insights into root nodulation in this species. *The Plant Journal*, **84**, 1–19. https://doi.org/10.1111/tpj.12967

Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, Kitzman J, Wu YQ, Newsham I, Richmond TA, Jeddeloh JA, Muzny D, Albert TJ, Gibbs RA (2010) Whole exome capture in solution with 3 Gbp of data. *Genome Biology*, **11**. https://doi.org/10.1186/gb-2010-11-6-r62

Burstin J, Avia K, Carillo-Perdomo E, Lecomte C, Beji S, Hanocq E, Aubert G, Tayeh N, Klein A, Geffroy V, Le Signor C, Pflieger S, Dalmais M, Desgroux A, Lavaud C, Quillévéré-Hamard A, Kreplak J, Lejeune-Hénaut I, Bourion V, Pilet-Nayel M, Leveugle M, Pinochet X, Thompson R, the PeaMUST Consortium (2021) PeaMUST (Pea MultiStress Tolerance), a multidisciplinary French project uniting researchers, plant breeders, and the food industry. *Legume Science*, **3**. https://doi.org/10.1002/leg3.108

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**. https://doi.org/10.4161/fly.19695

Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R (1998) Plant genome size estimation by flow cytometry: Inter-laboratory comparison. *Annals of Botany*, **82**. https://doi.org/10.1006/anbo.1998.0730

Duarte J, Rivière N, Baranger A, Aubert G, Burstin J, Cornet L, Lavaud C, Lejeune-Hénaut I, Martinant JP, Pichon JP, Pilet-Nayel ML, Boutet G (2014) Transcriptome sequencing for high throughput SNP

development and genetic mapping in Pea. *BMC Genomics*, **15**. https://doi.org/10.1186/1471-2164-15-126

Hill CB, Wong D, Tibbits J, Forrest K, Hayden M, Zhang XQ, Westcott S, Angessa TT, Li C (2019) Targeted enrichment by solution-based hybrid capture to identify genetic sequence variants in barley. *Scientific Data*, **6**. https://doi.org/10.1038/s41597-019-0011-z

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94. https://doi.org/10.1186/1471-2156-11-94

Kreplak J, Madoui MA, Cápal P, Novák P, Labadie K, Aubert G, Bayer PE, Gali KK, Syme RA, Main D, Klein A, Bérard A, Vrbová I, Fournier C, d'Agata L, Belser C, Berrabah W, Toegelová H, Milec Z, Vrána J, Lee HT, Kougbeadjo A, Térézol M, Huneau C, Turo CJ, Mohellibi N, Neumann P, Falque M, Gallardo K, McGee R, Tar'an B, Bendahmane A, Aury JM, Batley J, Le Paslier MC, Ellis N, Warkentin TD, Coyne CJ, Salse J, Edwards D, Lichtenzveig J, Macas J, Doležel J, Wincker P, Burstin J (2019) A reference genome for pea provides insight into legume genome evolution. *Nature Genetics*, **51**. https://doi.org/10.1038/s41588-019-0480-1

Kreplak J, Aubert G, Leveugle M, Duborjal H, Jean-Philippe Pichon, Burstin J (2021) Exome capture genotyping data on the pea Architecture Multi-Stress collection (Peamust project). Recherche Data Gouv, V4. https://doi.org/10.15454/3QRIPA

Ku CS, Wu M, Cooper DN, Naidoo N, Pawitan Y, Pang B, Iacopetta B, Soong R (2012) Exome versus transcriptome sequencing in identifying coding region variants. *Expert Review of Molecular Diagnostics*, **12**. https://doi.org/10.1586/erm.12.10

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**. https://doi.org/10.14806/ej.17.1.200

Mendel G (1866) Versuche über Pflanzenhybriden. In: *Versuche über Pflanzenhybriden* . https://doi.org/10.1007/978-3-663-19714-0_4

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era (E Teeling, Ed,). *Molecular Biology and Evolution*, **37**, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nawae W (2023) The value of a large Pisum SNP dataset. *Peer Community in Genomics*, **1**, 100237. https://doi.org/10.24072/pci.genomics.100237

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, **42**. https://doi.org/10.1038/ng.499

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**. https://doi.org/10.1038/nature08250

Ollivier R, Glory I, Cloteau R, Le Gallic JF, Denis G, Morlière S, Miteul H, Rivière JP, Lesné A, Klein A, Aubert G, Kreplak J, Burstin J, Pilet-Nayel ML, Simon JC, Sugio A (2022) A major-effect genetic locus, ApRVII, controlling resistance against both adapted and non-adapted aphid biotypes in pea. *Theoretical and Applied Genetics*. https://doi.org/10.1007/s00122-022-04050-x

Raj A, Stephens M, Pritchard JK (2014) FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, **197**. https://doi.org/10.1534/genetics.114.164350

Siol M, Jacquin F, Chabert-Martinello M, Smýkal P, Le Paslier MC, Aubert G, Burstin J (2017) Patterns of genetic structure and linkage disequilibrium in a large collection of pea germplasm. *G3: Genes, Genomes, Genetics*, **7**. https://doi.org/10.1534/g3.117.043471

Wang Y, Huang J, Li E, Xu S, Zhan Z, Zhang X, Yang Z, Guo F, Liu K, Liu D, others (2022) Phylogenomics and biogeography of Populus based on comprehensive sampling reveal deep-level relationships and multiple intercontinental dispersals. *Frontiers in Plant Science*, **13**, 8. https://doi.org/10.3389/fpls.2022.813177

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y (2017) ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, **8**, 28–36. https://doi.org/10.1111/2041-210X.12628

Zhang W, Dasmahapatra KK, Mallet J, Moreira GR, Kronforst MR (2016) Genome-wide introgression among distantly related Heliconius butterfly species. *Genome biology*, **17**, 1–15. https://doi.org/10.1186/s13059-016-0889-0