https://doi.org/10.35662/unine-thesis-3175

DETECTION AND CHARACTERIZATION OF EXOGENOUS DNA

From Genetically Modified Organisms (GMOs) to Naturally Admixed Genomes

A dissertation submitted to the University of Neuchâtel, Switzerland for the degree of Docteur ès sciences by

C. Sarai Reyes-Avila

Submitted to the Jury:

Prof. Daniel Croll, Université de Neuchâtel

Prof. Thomas Flatt, Université de Fribourg

Dr Sylvain Aubry, Federal Office for Agriculture

Prof. Daniel Wegmann, Université de Fribourg

Prof. Lothar Kalmbach, Université de Neuchâtel

unine
Université de Neuchâtel
Faculté des sciences

Rue Emile-Argand 11
CH-2000 Neuchâtel
doctorat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Madame C. Sarai REYES AVILA

Titre :

"Detection and characterization of exogenous DNA - from genetically modified organisms (GMOs) to naturally admixed genomes"

sur le rapport des membres du jury composé comme suit :

- **Prof. Daniel Croll,** directeur de thèse, Université de Neuchâtel, Suisse
- **Prof. ass. Lothar Kalmbach,** Université de Neuchâtel, Suisse
- **Prof. Thomas Flatt,** Université de Fribourg, Suisse
- **Prof. Daniel Wegmann,** Université de Fribourg, Suisse
- **Dr Sylvain Aubry,** OFAG, Suisse

Neuchâtel, le 1er avril 2025

Le Doyen, Prof. P. Brunner

TABLE OF CONTENTS

SUMMARY

This thesis focuses on detecting exogenous DNA, whether it arises naturally through admixture or is introduced through genetic modification, by advancing methods for identifying these genetic traces. In Chapter 1, we developed a highly multiplexed amplicon sequencing assay to detect first-generation GMOs. Our assays use a microfluidics platform and next-generation sequencing (NGS) to amplify in parallel and sequence multiple GMO targets. We designed 230 primer pairs to amplify GM events across different crops. We also included barcoding markers for species identification. We demonstrated that our assay can detect first-geneation GMOs in a parallel amplification. Our assay also uncovered potential “unknown” GM events that standard PCR screens might miss. Given its scalability in simultaneously processing multiple targets and samples, our microfluidics-based assay can serve as a first-pass screening tool. It enables broad detection that can be reviewed with confirmatory methods. In Chapter 2, we introduced LOCO (LOw depth COpy algorithm), a new computational model to infer local ancestry from low-coverage sequencing data without depending on external reference panels. LOCO builds upon Li & Stephens–style copy models but constructs its reference haplotypes directly from the data. By simulation tests, we demostrated that LOCO can infer the correct ancestry in admixed genomes and detect longer introgressions. However, we observed that short ancestry segments are often misassigned, a common limitation of local-ancestry tools. In Chapter 3, we applied this ancestry-inference idea to second-generation GMO detection. Here, we simulated second-generation GMO-like modification in rice genomes by artificially copying small segments from one individual into another. In principle, LOCO should flag these segments as introgression if they are different from the individual’s ancestry background. However, we encountered difficulties with initialising the parameters needed by LOCO. Given that the set of parameters used by LOCO are unknown for this data set, LOCO failed to find the global maxima solutions. We need a better strategy for initialising the parameters from real-world data bceause we rarely know the true value of the parameters. Overall, this thesis demonstrates thats our sequencing and computational methods can significantly improve the detection of exogeneus DNA.

RESUME

Cette thèse porte sur la détection de l'ADN exogène, qu'il provienne naturellement par métissage ou qu'il soit introduit par modification génétique, en développant des méthodes permettant d'identifier ces traces génétiques. Au Chapitre 1, nous avons développé un test de séquençage d'amplicons hautement multiplexé pour détecter les OGM de première génération. Nos tests utilisent une plateforme de microfluidique et le séquençage de nouvelle génération (NGS) pour amplifier en parallèle et séquencer plusieurs cibles OGM. Nous avons conçu 230 paires d'amorces pour amplifier des événements de modification génétique dans différentes cultures. Nous avons également inclus des marqueurs de codage-barres pour l'identification des espèces. Nous avons démontré que notre test peut détecter les OGM de première génération par amplification parallèle. Notre test a également révélé des événements OGM « inconnus » que les tests PCR standards pourraient ne pas détecter. Étant donné sa capacité à traiter simultanément plusieurs cibles et échantillons, notre méthode basée sur la microfluidique peut servir d'outil de dépistage initial. Elle permet une détection large qui peut ensuite être examinée à l'aide de méthodes de confirmation plus sensibles. Au Chapitre 2, nous avons présenté LOCO (algorithme de COpy à faible profondeur), un nouveau modèle computationnel permettant d'inférer l'ascendance locale à partir de données de séquençage à faible couverture, sans dépendre de panels de référence externes. LOCO s'appuie sur des modèles de type Li & Stephens, mais construit ses haplotypes de référence directement à partir des données. À travers des simulations, nous avons démontré que LOCO peut inférer correctement l'ascendance dans des génomes issus d'admixture et détecter de longues introgressions. Toutefois, nous avons observé que les segments courts d'ascendance sont souvent mal attribués, une limitation fréquente des outils d'inférence d'ascendance locale. Au Chapitre 3, nous avons appliqué cette approche d'inférence d'ascendance à la détection d'OGM de seconde génération. Nous avons simulé, dans ce cadre, une modification de type OGM de seconde génération dans des génomes de riz, en copiant artificiellement de petits segments d'un individu à un autre. En principe, LOCO devrait identifier ces segments comme des introgressions s'ils diffèrent du fond d'ascendance de l'individu. Toutefois, nous avons rencontré des difficultés à initialiser les paramètres nécessaires au bon fonctionnement de LOCO. Étant donné que l'ensemble des paramètres requis est inconnu pour cet ensemble de données, LOCO n'a pas réussi à trouver les solutions de maximum global. Il est donc nécessaire de développer une meilleure stratégie d'initialisation des paramètres pour les données réelles, car dans la pratique, la vraie valeur de ces paramètres est rarement connue. Dans l'ensemble, cette thèse démontre que nos méthodes de séquençage et nos outils computationnels peuvent considérablement améliorer la détection de l'ADN exogène.

GENERAL INTRODUCTION

Exogenous DNA can be defined as a genetic piece that comes from outside an organism and integrates into the genome of that organism. This integration can happen naturally or by human-made introgression. Hybridisation is a type of natural introgression when individuals from different species reproduce, and their progeny inherit the genetic material from both parents (Baack & Rieseberg, 2007; Schwenk et al., 2008). In humans, admixture, the process where genetic material is exchanged between populations, represents a natural way of generating exogenous DNA. The genome of an admixed individual is a mixture formed from two or more diverged populations (Korunes & Goldberg, 2021). Each chromosome is a mosaic of segments derived from a particular ancestral population. Admixture shapes human genetic diversity, influences traits and contributes to disease susceptibility (Divers et al., 2017; Garzón Rodríguez et al., 2024; Gomez et al., 2015; Koller et al., 2022; Pankratov et al., 2024; Welter et al., 2014; Xia et al., 2024). Genome-wide association Studies (GWAS) have identified more than one thousand loci, specific positions of the genome, in the human genome associated with complex traits (Welter et al., 2014). Combining GWAS and methods for admixture detection helps to understand how different ancestral backgrounds contribute to certain diseases. For example, the European biobank found that post-Neolithic admixture events contributed to genetic predisposition traits related to heart rate and bone mineral density (Pankratov et al., 2024). Other studies of African Americans demonstrated that admixture mapping can identify loci related to coronary artery calcification, which is an indicator of atherosclerosis, and that loci differ between African and European ancestry (Divers et al., 2017; Gomez et al., 2015). Another study in Latin American populations identified specific genetic variants at loci associated with increased risk for ADHD that were predominantly inherited from European ancestry (Garzón Rodríguez et al., 2024). In another study, non-human but hominin Denisovan loci related to coronary artery disease heritability were found in East Asian populations (Koller et al., 2022). Neanderthal alleles were associated with traits like red hair colour and psychiatric conditions in Europeans (Koller et al., 2022). A study about the Major Histocompatibility Complex (MHC), which helps the body recognise foreign substances,

found that conserved ancestral haplotypes, which have been passed over many generations, carry susceptibility and resistance to autoimmune diseases (Dawkins & Lloyd, 2019). In another non-human study in grapevine*, Vitis vinifera,* to investigate the role of historical admixture in shaping genetic diversity, it was found that domesticated grapevines have signatures of admixture from wild grape populations, which influenced traits related to fruit size, sugar content, and disease resistance. This study highlights how admixture mapping can be used in plants to identify genomic regions associated with desirable traits. All these studies demonstrate the relevance of studying admixture because it helps to uncover ancestry-specific genetic risk factors across different populations.

On the other hand, a human-made introgression involves the deliberate introduction of exogenous DNA into an organism's genome using biotechnological techniques for genetic modification. Early techniques started with untargeted mutagenesis, using chemical agents or radiation to induce random mutations that lacked precision (*The Production of Mutations by X-Rays - PMC*, n.d.). Another group of untargeted techniques use vectors, DNA molecules that carry and introduce genetic material into cells. *Agrobacterium tumefaciens* can transfer DNA into plant genomes, and this skill is used as a tool for genetic transformation in plants (Zambryski et al., 1983). Other transformations can be done chemically with calcium phosphate or polyethene glycol to facilitate DNA ingestion, an electrical transformation that uses electric pulses to create membrane pores for DNA entry, and particle bombardment where DNA microparticles are physically sent into cells using high-velocity propulsion (“Delivery of Substances into Cells and Tissues Using a Particle Bombardment Process,” 1987; *Gene Transfer into Mouse Lyoma Cells by Electroporation in High Electric Fields - PubMed*, n.d.; Ozyigit, 2020). Genetic engineering techniques continue improving, nowadays, we have tools that allow for targeted editing. These targeted tools are Zinc-Finger Nucleases (ZFNs), Transcription Activator-Like Effector Nucleases (TALENs), and the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas9. ZFNs are synthetic proteins with a zinc finger that acts as a DNA-binding domain, these zinc fingers act together with a nuclease to create double-strand breaks at a specific location. This is a targeted approach that can break the DNA, and the natural cell's mechanisms will repair it; if

there is a donor DNA with homology, it will be introduced to the genome (Gaj et al., 2013). However, the design and assembly of ZFNs for each target sequence can be complex and time-consuming (Gaj et al., 2013). TALENs use transcription activator-like effectors, which recognize specific DNA sequences, and combined with a nuclease will induce the break in the DNA (Gupta & Musunuru, 2014). TALENs are easier to design than ZFNs, but they still need customized proteins for each target. The CRISPR-Cas9 system revolutionised the field of genome editing, given its simplicity and versatility. CRISPR-Cas9 has a single-guide RNA (sgRNA) to direct the Cas9 nuclease to the target where it breaks the DNA; this allows precise targeted genetic modifications, and it is cost-effective (Gaj et al., 2013; Gupta & Musunuru, 2014).

First-generation genetically modified organisms (GMOs) are developed using genetic engineering methods that rely on inserting large exogenous DNA constructs into the host genomes. These constructs typically contain promoters, terminators, coding sequences, and regulatory elements. When a construct is successfully integrated into the genome of an organism, the resulting integration site is referred to as a GMO event. These events are randomly integrated using methods like Agrobacterium-mediated transformation (Arulandhu et al., 2018; Scholtens et al., 2017). There are numerous examples of GMOs that have been created to give traits to plants, such as herbicide tolerance, insect resistance, and tolerance to abiotic stress. For example, soybean has been genetically modified to express a gene encoding 5-enolpyruvylshikimate-3-phosphate synthase, which provides resistance to glyphosate-based herbicides. The gene originates from an *Agrobacterium sp.* strain CP4, which naturally resists glyphosate. For introducing the gene, *A. tumefaciens* was used for plant transformation and transferred the gene to the plant genome (Arias et al., 2021). For the case of insect resistance, an example is corn that was engineered to express the Cry1Ab protein from *Bacillus thuringiensis* to protect against pests like the European corn borer (Priestley & Brownbridge, 2009). Another example is canola, which was engineered to have an altered fatty acid composition, enhancing its industrial applications (Neff et al., 1994). In addition to these traits, plants have also been modified to improve nutritional content. For example, “golden rice” was engineered to produce provitamin A to address vitamin A deficiencies in regions where the human diet

depends on rice as a principal food source (Ye et al., 2000). A similar modification produced the "golden banana" (Paul et al., 2017).

In contrast to inserting large DNA constructs into the organism's "first-generation GMOs", the New Genomic Techniques (NGTs), specifically CRISPR-Cas9, allow site-specific genome modifications. This CRISPR-Cas9 system uses a single-guide RNA (sgRNA) to direct the Cas9 nuclease to a target DNA sequence, which induces a double-strand break. The cell naturally repairs the non-homologous end joining or via homology-directed repair when a donor template is provided. This allows for precise nucleotide substitutions, gene knockouts, or small insertions/deletions (Gaj et al., 2013; Gupta & Musunuru, 2014). The RNA-guided modularity of CRISPR-Cas9 allows a simple reprogramming by altering the sgRNA sequence. CRISPR-Cas9 has a high level of specificity and reduces off-target effects compared to earlier techniques. The generation of site-specific genome modifications is called second-generation GMOs. The type of genetic modification, whether a single nucleotide modification or the insertion of a large transgenic construct, has different implications. Nucleotide-level edits have lower risks of off-target effects (Hsu et al., n.d.; Sander & Joung, 2014). These modifications do not involve the integration of large exogenous DNA fragments that can disrupt regulatory elements or chromatin organisation (Hsu et al., n.d.; Sander & Joung, 2014). In contrast, large insertions from first-generation GMOs, which include multiple regulatory elements like promoters, terminators, and coding sequences in their constructs, might interfere with endogenous gene expression. Also, the unpredictable integration sites and potential off-target disruptions need more screening (Sander & Joung, 2014). Agricultural applications include editing hexaploid bread wheat with CRISPR/Cas9 at all three homoeo-alleles of the MLO gene, conferring broad-spectrum resistance to the powdery mildew pathogen (Y. Wang et al., 2014). In rice, CRISPR/Cas9 was used to edit the amino acid transporter gene *OsAAP3,* which enhances grain yield by promoting the outgrowth of buds (Lu et al., 2018). In tomatoes, CRISPR/Cas9 was used to knock out the genes SlINVINH1 and SlVPE5, increasing the soluble sugar content (B. Wang et al., 2021). Furthermore, CRISPR/Cas9 was used to edit specific carotenoid biosynthesis genes in tomatoes to enhance the lycopene content and red colour of fruits (Tiwari et al., 2023). In bananas, CRISPR/Cas9 was

used to inactivate the endogenous banana streak virus (eBSV) to prevent the activation of infectious viral particles (Tripathi et al., 2019). The development of GM rice (*Oryza sativa*) varieties highlights the adoption of CRISPR/Cas9, representing 12-33% of global gene-editing research compared to other crops (Chen et al., 2024). For example, Wang *et al.* knocked out the Pi21 gene to produce rice varieties with resistance to rice blast (F. Wang et al., 2016). Zeng *et al.* edited OsSWEET14 to confer resistance against bacterial blight caused by *Xanthomonas oryzae* (Zeng et al., 2020). Lu *et al.* knocked out OsSPL10 to enhance rice's resistance to the brown planthopper, a major pest (Lu et al., 2018). Zhang *et al.* edited OsRR22 to develop rice varieties with enhanced salt tolerance (Zhang et al., 2019). Zhao *et al.* edited the GS9 gene to increase grain length (Zhao et al., 2018). Zeng *et al.* edited the Wx promoter region, which influences the texture and taste of cooked rice (Zeng et al., 2020).

Given these advancements in agriculture, the regulation of GMOs also has been progressing worldwide, focusing on safety and transparency in the development of GMOs and their uses. One of the first efforts was established in 1986 with “The Coordinated Framework for the Regulation of Biotechnology”, which describes the roles of the U.S. Food and Drug Administration (FDA), Environmental Protection Agency (EPA), and U.S. Department of Agriculture (USDA) in the supervision of GMOs (Program, 2024). In general, GMO regulations evaluate GMOs to estimate potential risks to human health and the environment. It involves safety assessments, like allergenicity and toxicity studies, and environmental impact evaluations. GMO regulation covers detection that involves identifying and quantifying GMOs in products and the environment. Nowadays, discrepancies exist in the global regulatory frameworks governing NGTs. In the U.S., gene-edited crops that do not contain exogeneous DNA are exempt from the rigorous regulatory process imposed on first-generation GMOs (Salt, 2023). The European Union (EU) initially classified all gene-edited organisms as GMOs, subjecting them to the same regulations (Schmidt et al., 2020). Directive 2001/18/EC is the primary legislation regulating the release of GMOs into the environment in the EU. The purpose of this directive was to regulate GMOs developed by using earlier genetic engineering techniques, first-generation GMOs. Advances in NGTs resulted in genetic modifications that may be indistinguishable from those

occurring naturally, second-generation GMOs. These advances generated concerns about whether the existing legal framework is suitable for regulating second-generation GMOs. (New Genomic Techniques, 2024). The revision process of Directive 2001/18 is ongoing; it aims to refine the regulatory framework for gene-edited crops (New Genomic Techniques, 2024). While genetic engineering continues advancing, the legal framework continues adapting, and the outcomes of these revisions remain to be seen.

EU regulations include specific provisions for detecting and quantifying GMO material in food and feed products (*GMOs - European Commission*, n.d.). Detection refers to identifying the presence of any GMO material in a product (Holst-Jensen et al., 2016). Quantification measures how much GMO content is present (Aubry et al., 2021). Quantification is necessary to determine if the GMO content exceeds labelling thresholds (*GMOs - European Commission*, n.d.). A positive GMO test in food/feed is followed by the quantification of the GMO percentage to check compliance with labelling rules. For food and feed products, there are authorised GMOs, those approved for use in the EU; the threshold per ingredient is 0.9% of GMO presence (*GMOs - European Commission*, n.d.). If an ingredient contains GMO material at 0.9% or above, the final product must be labelled as “contains GMO”. EU regulatory compliance testing relies on robust analytical methods to detect and quantify GMOs. In the case of the detection and quantification of first-generation GMOs, there is a broad range of methods, but still, quantitative polymerase chain reaction (qPCR) approaches remain the standard technique given their sensitivity and specificity (Broeders et al., 2012; Marmiroli et al., 2008). qPCRs allow the detection, identification, and quantification of genetic modifications based on PCR amplification recorded in real time. As an example, a semiautomated TaqMan PCR screening was developed and tested based on 32 primer and probe sequences derived from methods used in routine GMO feed sample screenings (Scholtens et al., 2017). Multiplex PCR-based methods were developed to facilitate the detection of several DNA targets in a single reaction. The challenges are to ensure that all targets are amplified efficiently and specifically within a multiplex reaction, given differential amplification efficiencies and the risk of primer-dimer formation (Dragoni et al., 2021; Lehnert & M. Gijs, 2024; Scholtens et al., 2017). Next-generation sequencing (NGS) appeared as a powerful tool for enabling the

characterization of multiple genetic sequences (Arulandhu et al., 2018). Next-generation sequencing (NGS) can help detect first-generation GMOs, determine insertion locations, copy number variations, and possible off-target mutations. For example, Arulandhu *et al.* use NGS-based screening to detect first-generation GMOs. Arulandhu's method can detect additional GM targets that are not detected with traditional qPCR by identifying low-abundance genetic elements (Arulandhu et al., 2018). The high resolution of NGS can be useful for the detection of genome edits that are challenging to identify using traditional PCR-based methods (Grohmann et al., 2019). The sequences obtained by NGS need to be mapped to reference databases (Arulandhu et al., 2018).GMO detection is a challenge, given the incomplete public disclosure of GM sequences (Aubry et al., 2021; Saltykova et al., 2022). If primers to detect a GMO were not developed for routine screening, the GMO is called an unknown (*i.e.* unauthorised) GMO. PCR methods share the limitation of requiring prior knowledge of the genetic modification to design specific primers, which limits the identification of unknown GMOs. The use of microfluidic technologies, which manipulate small volumes of fluids within microscale channels and compartments to make their reactions, offers advancements for applications requiring high throughput and rapid diagnostics (Dragoni et al., 2021; Lehnert & M. Gijs, 2024). Microfluidics reduces multiple reaction times, lowers reagent volumes, and increases throughput by simultaneous processing of multiple samples. Microfluidics have shown high sensitivity and faster processing, potentially being more suitable for GMO detection compared to traditional qPCR methods (Dragoni et al., 2021).

In the case of detection for the second generation of GMOs, PCR-based techniques are also the main method used (Chhalliyil et al., 2020). In canola, a single-nucleotide modification has been quantified using a qPCR for the first commercially available genome-edited plant (Chhalliyil et al., 2020). The Joint Research Centre (JRC), which is the European Union Reference Laboratory for Genetically Modified Food and Feed, applies these qPCR methods and has been involved in their development (Chhalliyil et al., 2020). However, PCR continues to have the limitation of prior knowledge of gene editing to design specific primers (Broeders et al., 2012). Also, the use of qPCR is problematic for precisely detecting edits introduced by NGT because the second

generation of GMOs involves single nucleotide modifications that are difficult to detect by standard qPCR assays (Aubry et al., 2021). Whole-genome sequencing (WGS) can help to detect modifications without prior knowledge of the modified sequence. This can be a solution for genome editing where no information on the sequence is available (Grohmann et al., 2019). The process is mapping sequencing reads from the sample to a reference genome sequence and identifying the differences in the sample, revealing potential genome edits. However, if the modification originates from a closely related species, it would be difficult to distinguish between the product of genome editing and natural sequence variants. In this scenario, the sequence similarity may suggest that the changes are the result of natural genetic variation rather than genome editing. Given this complexity, there is a strong demand to improve detection procedures, and bioinformatics analysis and development of powerful statistical algorithms addressing this challenge.

The genetic exchange of individuals from the same species results in the natural introduction of exogenous DNA from one individual into another, and this process is called admixture. In diploid organisms, where each individual inherits two sets of chromosomes, admixture produces a genome that is a mosaic of segments derived from two or more diverged populations. (Korunes & Goldberg, 2021). These genomics segments originate from recombination events during meiosis, where homologous chromosome pairs exchange DNA segments through a crossing-over mechanism (Kleckner, 2006). The crossing-over happens in specific sites called chiasmata, the point of physical contact between two chromatids belonging to the homologous chromosomes. There, physical breakage and rejoining lead to the reshuffling of alleles (Kleckner, 2006; Wegmann et al.., 2011). As a result, “switch points” are created where the ancestral origin shifts from one population to another (Wegmann et al., 2011). In general terms, to characterize admixture, we need to 1) determine whether an individual is admixed, 2) map the genomic segments inherited from each ancestral population by identifying recombination switch points, and 3) assign these segments to their respective populations using reference panels (Salter-Townshend & Myers, 2019). Reference panels are curated genotype data from individuals with well-characterised ancestry (Salter-Townshend & Myers, 2019). When determining the ancestral origins of

genomic regions in admixed individuals, the reference panels help to identify switch sites and determine ancestry proportions by comparing the genetic markers in an admixed genome to those in the reference panel (Mosca & Cho, 2023; Salter-Townshend & Myers, 2019). For example, Bryc *et al.*. used reference panels from West African and European individuals to estimate the ancestry proportions in African-American populations (Bryc et al., 2010). Their analysis showed a mosaic genome structure, where segments could be confidently assigned to West African or European origins (Bryc et al., 2010).

The assignment of ancestry for the admixed individuals requires probabilistic methods because different ancestral populations often shared genetic markers, given their common evolutionary history (Shriner, 2013). Methods for ancestry inference are broadly categorized into allele-frequency-based and haplotype-based models (Padhukasahasram, 2014). Allele-frequency-based methods focus on how common alleles are in the reference populations. This approach was implemented in the LAMP (Local Ancestry in adMixed Populations) model, which determines the most likely ancestry at each locus using a majority voting system (Sankararaman et al., 2008). LAMP requires no reference panel, which is an advantage in cases where ancestral populations of interest cannot be adequately sampled or when representative genetic data is unavailable. LAMP is highly accurate in differentiating between well-separated populations (for example, Africans and Europeans) compared to methods that rely on haplotype information. However, it does not perform well when the ancestral populations are very similar (Yuan et al., 2017). STRUCTURE and ADMIXTURE are known allele-frequency-based tools. Each one of them approaches the problem with different algorithms. STRUCTURE employs a Bayesian framework to assign individuals to predefined populations and infer population structure without prior knowledge of population boundaries (Pritchard et al., 2000). ADMIXTURE is a maximum-likelihood-based method optimized for speed and scalability, making it useful for genome-wide datasets with thousands of samples (Alexander et al., 2009).

On the other hand, haplotype-based methods use Hidden Markov Models (HMMs) to leverage haplotype structure for ancestry at a finer scale compared to methods that only

rely on allele frequencies (Price et al., 2009; Shriner, 2013). This finer resolution comes when haplotype-based methods consider the contiguous segments of DNA inherited together from ancestors, allowing them to detect smaller, more localised ancestry contributions than allele-frequency-based methods (Guan, 2014). This approach was implemented in HAPMIX (HAPlotype MIXture) (Price et al., 2009), which uses phased reference panels to model the ancestry process of assignment; this allowed the detection of small genomic segments that have distinct ancestral origins. HAPMIX relies on phased reference data to estimate the likelihood that a given haplotype segment originates from one population or another and also relies on detailed recombination maps, which limits its applicability, especially for populations with insufficiently characterized or unavailable reference panels data. (Price et al., 2009). Methods such as RASPberry have been developed to eliminate the dependency on *a priori* established recombination maps. RASPberry analyzes patterns of genetic variation and ancestry transitions in admixed individuals to infer recombination rates directly from genotype data (Wegmann et al., 2011). Regardless of the progress in ancestry inference made by the implementation of HMM models, challenges persist by not being able to directly model sequencing data and estimate reference panels from the data. Both producing high-quality sequencing data -with sufficient coverage and low error rates- and establishing reference panels are costly and often unattainable for non-human species or ancient DNA applications (Shriner, 2013).

Admixture analysis relies on the availability of high-quality sequencing data and accurate reference panels representing the ancestral populations of interest. In many genomic studies, data quality can be affected by factors such as low sequencing coverage, genotyping errors, or missing data due to the limitations of genotyping platforms (Browning & Browning, 2016). These problems lead to incomplete genotype datasets. The genotype imputation field focuses on predicting missing genotypes. It uses the linkage disequilibrium structure in a population, defined as the non-random association of alleles at different loci (Browning & Browning, 2016; Howie et al., 2012). Genotype imputation uses well-characterized reference panels as well as the admixture analysis previously described. Imputation methods rely on probabilistic frameworks to infer missing genotype data from observed patterns (Howie et al., 2012). For example,

IMPUTE2 is a tool based on an HMM framework that treats the unobserved true haplotypes as hidden states (Howie et al., 2012). This HMM generates the observed genotype data from these hidden states, representing the underlying haplotype structure from the well-characterized reference panel. This approach allows IMPUTE2 to accurately predict missing genotypes by "copying" segments from the reference haplotypes, thereby maintaining the linkage disequilibrium patterns present in the data (Howie et al., 2012). The advantage of this approach is its high imputation accuracy when a reference panel is available. However, dependence on external reference panels can also be a limitation for populations that lack well-curated reference panels. Another tool is BEAGLE, which uses localised haplotype clustering rather than a global HMM framework (Browning & Browning, 2016). BEAGLE identifies clusters of similar haplotypes in small genomic regions by capturing the local linkage disequilibrium structure. It uses these clusters to predict missing genotypes based on the observed alleles within them. This localised approach makes BEAGLE more computationally efficient than IMPUTE2 and is not dependent on reference panels but only on its available data. However, the limitation is that BEAGLE may offer lower accuracy compared to IMPUTE2 in scenarios when fine-scale haplotype structure is critical, for example, when subtle differences in the arrangement of alleles along a chromosome are important for accurately predicting missing genotypes. Both IMPUTE2 and BEAGLE perform best when genotype data have high call rates, low error rates, and sufficient marker density to capture linkage disequilibrium patterns across the genome. STITCH is an alternative tool for cases where high-quality data and reference panels are limited (Davies et al., 2016). STITCH (Sequencing To Imputation Through Constructing Haplotypes) addresses the challenges derived from low-coverage sequencing data by directly modelling the genotypes while accounting for the for the uncertainty inherent in sequencing reads. STITCH uses raw sequencing data to estimate the probability of various genotypes, equivalent to genotype likelihood information. Genotype likelihoods quantify the probability of obtaining the observed sequencing data given a particular genotype and are used to measure confidence in the genotype call; here, uncertainty is kept rather than forcing a definitive call. By using an expectation-maximization (EM) algorithm, STITCH iteratively estimates the haplotype structure and the missing

genotypes from low-coverage data without requiring external reference panels. As a result, STITCH can generate imputed genotypes with quality scores that reflect the confidence of the inference, all without the need for external reference panels. It is important to note that STITCH does effective genotype imputation under low sequencing coverage, but STITCH does not perform detailed ancestry inference. STITCH lacks features critical for ancestry inference, such as explicitly modelling ancestry transitions, recombination events, or miscopying errors, which are essential for understanding the mosaic structure of admixed genomes.

THE RATIONALE OF THE THESIS

Regardless of the significant advancements in detecting exogenous DNA, certain gaps persist. 1) Current detection methods like PCR and qPCR are well-established for identifying first-generation GMOs. However, these techniques have scalability issues, particularly for high throughput. Traditional PCR-based assays are often limited in their multiplexing capabilities. Currently, the maximum number of samples that can be tested at once is limited. 2) Ancestry inference methods are important for understanding genetic diversity in populations. However, these methods often require high-quality sequencing data and well-characterized reference panels, which are not always available, especially for underrepresented populations. 3) Traditional PCR techniques can detect large transgenic insertions but are not optimal for detecting minimal gene edits of the second generation of GMOs because these subtle modifications can be indistinguishable from natural genetic variations.

The primary objective of this thesis is to develop and test new methodologies for the detection and characterization of exogenous DNA, focusing on natural and human-made exogenous DNA. This thesis contains three main chapters that address the primary objective in the following way:

Chapter 1: Detection of Genetically Modified Organisms Using Highly Multiplexed Amplicon Sequencing

In this chapter, we developed a highly multiplexed amplicon sequencing assay that also uses NGS.We collected a set of GMO reference sequences to design amplicon targets for the genetic modifications. The assay is evaluated in terms of specificity and sensitivity across various samples.

Chapter 2: LOCO: Low-Depth Copy Algorithm to Infer the Ancestry of an Admixed Individual Without Reference Panels

In this chapter, we developed a novel probabilistic model to infer ancestry without relying on reference panels and using genotype likelihood derived from low-quality sequencing data. We established the mathematical formulation of the model, its implementation in C++, and validation by defining a set of parameters that we use to produce simulated data. With this data, we recover the ancestry of the simulation. Furthermore, we simulate introgression events and apply the LOCO model to detect these events, evaluating the model's detection limits.

Chapter 3: Machine Learning for Detection of Next-Generation Genetically Modified Organisms Through Ancestry Inference

In the final chapter, we applied our model for ancestry inference from Chapter 2 to identify introgressions associated with second-generation GMOs. We simulate genome edits in a rice population to create a set of second-generation GMOs and evaluate the detection power of our LOCO model. Our simulations replicate the scenario where introgression occurs between two closely related populations within the same species.

CHAPTER 1

Detection Of Genetically Modified Organisms Using Highly Multiplexed Amplicon Sequencing

C. Sarai Reyes-Avila[1], Dominique Waldvogel[2], Nicolas Pradervand[3], Sylvain Aubry[4,5], Daniel Croll[1,*]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, CH-2000 Neuchâtel, Switzerland

[2] Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland

[3] Agroscope. Posieux, Switzerland

[4] Department of Plant and Microbial Biology, University of Zürich, Zürich, Switzerland

[5] Federal Office for Agriculture, Bern, Switzerland

[*] Correspondence: daniel.croll@unine.ch

Keywords: genetically modified organisms, crops, multiplex PCR, NGS, amplicon-sequencing

ABSTRACT: The circulation of products based on genetically modified (GM) organisms is highly regulated by some governments with strict implementation rules for the breeding, planting, marketing, labelling, and trading of such products. To ensure compliance, accurate detection methods for GM events are necessary, along with assurance that GM material falls within relevant threshold levels. The increasing complexity and potential of undocumented GM are a growing challenge for genetic screening. Here, we developed and assessed a highly multiplexed amplicon sequencing assay for the detection of GM events based on a microfluidics platform and next-generation sequencing (NGS). To probe GM events comprehensively, we designed a total of 230 new amplicons to cover flanking, promoter, junction, and coding sequences of GM sequences. In addition, we designed and implemented parallel amplification of ribosomal and chloroplast markers to define crop species identity from potentially mixed samples. Using reference GM material of 11 crop species and multiple amplicons, we successfully detected the presence of 10 known modifications per GM event. We also find that reported flanking sequences of GM events may not be all useful for diagnostic. We assessed the assay's potential to detect GM events in mixed samples as well as in highly diluted DNA. Finally, we performed a prospective search of potentially undocumented GM events in plant material. Our microfluidics-based amplicon GM detection approach fills important gaps in detecting potentially undocumented and complex GM events by recovering a wide range of specific amplicon sequences for evaluation. Integrating highly parallel amplicon assays in GM screening efforts should be an effective complement to aid post-market monitoring and regulatory compliance efforts.

INTRODUCTION

Genetically modified (GM) crops contain genes that have been artificially introduced conferring beneficial traits such as herbicide tolerance, and drought or pest resistance (Alasaad, Alzubi, and Kader 2016). The production of GM organisms has been rising and becoming more widely available commercially. Commercial applications of GM in crops have led to hundreds of variants of genetic modifications in dozens of crop species but mostly in the main cash crops (maize, soybean, canola and cotton). Since the ’90s, the European Union (EU) implemented risk-based legislation governing the planting, marketing, labelling and trade of GMOs in Europe (Serageldin 1999). To enforce legislation, it is essential to develop accurate methods for the detection of GM and monitoring of threshold levels.

Any material deriving from GM crops might be identified by testing for the presence of introduced DNA. GM events or constructs that are often found in commercialized GM

crops usually consist of several elements (promoters, terminators, genes, antibiotic resistance cassettes) that help to screen for their presence. There is a broad range of methods for GM DNA detection, and quantitative polymerase chain reaction (qPCR) approaches remain the most common (Aubry et al. 2021). qPCRs allow the detection, identification, and quantification of genetic modifications, based on PCR amplification recorded in real-time. As an example, a semiautomated TaqMan PCR screening of GMO-labelled samples was recently developed (Scholtens et al. 2017). Scholtens *et al.* semiautomated screening based on 32 primer and probe sequences derived from methods used in routine GMO feed sample screenings. Scholtens *et al.* verified the 32 primers in 59 different GMO reference materials. When a GMO element cannot be explained by any sample labelling, this is indicative of an unknown (*i.e.* unauthorised) GMO.

To facilitate the detection of several DNA targets in a single reaction, multiplex PCR-based methods were developed. qPCR-based multiplexing is constrained through costs and technical challenges to retain high reproducibility across laboratories. Official control laboratories are required certification to ensure high reproducibility standards. In addition, strategies have been developed to increase assay throughput without compromising sensitivity or specificity.

The introduction of microfluidic chips in multiplex PCR amplification offered significant gains in efficiency. Studies *e.g.* by Kuan-Lun et al. (2023), Nouwairi et al. (2022), Dong et al. (2021), Yang et al. (2021), and Lehnert & Gijs (2024) showed that reaction times can be reduced with lower reagent volume requirements and robust target multiplexing. Nouwairi et al. achieved reduced PCR cycle time with a custom microfluidic chips (Nouwairi et al. 2022). Yang et al. showcased a novel circular array-shaped microfluidic chip that allows high-throughput detection of bacterial pathogens, significantly decreasing the time and spatial requirements typical of conventional PCR setups (Yang et al. 2023). Lehnert & Gijs (2024) showed that microfluidic systems can be useful for rapid diagnostics in clinical settings (Lehnert and Gijs 2024). Microfludics offers the most powerful options in combination with next-generation sequencing (NGS) of amplicons. NGS has been used extensively for GM detection and allows to

simultaneously obtain multiple sequences and may facilitate the unambiguous detection of GMOs (Arulandhu et al. 2018). Arulandhu *et al.* tested five feed samples known to contain GMOs and screened 96 GMO-specific targets including endogenous, elements, constructs, and different events. Based on a related approach, Scholtens *et al.* targeted sequences from certified reference materials. Targeted sequencing was mainly applied to samples with multiple GM sequences of interest (Jagadeesan et al. 2019). DNA libraries were obtained from PCR-based amplification including several amplicons that can be sequenced using NGS technology. Standardized bioinformatics pipelines were applied to raw reads for filtering and assembly into contigs, which can be mapped to a reference sequence database (Willems et al. 2016). Establishing reference databases remains challenging as public disclosure of modified sequences can be incomplete (Moreira, Carneiro, and Pereira 2017a).

GMO detection traditionally was anchored in the simultaneous amplification of endogenous reference genes. Such genes typically amplify only in specific crop varieties where the GMO was initially developed, limiting how well an assay can be ported across diverse genetic backgrounds (Huang et al. 2013). Moreover, endogenous genes may vary significantly across different species, which reduces their utility for studies involving multiple species or agnostic assays. For example, in sugarcane, the selection of suitable endogenous reference genes is challenging due to the crop's complex polyploid and aneuploid genome structure. Such variability in traditional reference genes underscores the need for novel approaches. With the rapid increase in GMO testing needs, there is also a demand for standardization of GMO detection approaches to maintain versatility across the vast number of materials to be tested (Dong et al. 2008).

To address these challenges, we propose adopting a novel approach based on standarding DNA-barcoding loci typically used for phylogenetic reconstruction. Furthermore, we avoided host-exogenous DNA boundaries in primer design to allow for cross-amplification of the same GM sequence in different backgrounds. This design strategy ensures that our assay can universally detect GM events without being limited

by the host plant's genetic background, making it broadly applicable to various GM crops. To achieve these aims, we developed a highly multiplexed amplicon sequencing assay for the detection of GM events in crops. The is assays was conceived by integrating a large set of available GM sequences from public databases to design new, robust amplicons to amplify sequences linked to GM events in segments. Barcoding primers were designed to detect the species identity present in the material. Unlike traditional approaches that rely on endogenous reference genes as controls in GM crop detection, this study employs barcoded genes to identify the origin of tested samples. This innovative approach allows for the broad applicability of GMO detection across various plant species, overcoming the limitations of species-specific PCR which restricts analysis to detecting presence or absence of specific species. Amplifications were conducted in parallel using a microfluidics-based multiplex PCR and amplicons were sequenced as a pool using Illumina NGS. Filtered reads of positive and negative control samples were mapped to a compiled GM amplicon database. The performance of the assay, in particular specificity and sensitivity, in diluted and mixed samples were evaluated.

MATERIALS AND METHODS

Collection of samples

Genotyping was performed on 92 plant samples consisting primarily of GM-certified reference materials (Table S1). Samples covered eleven of the most widely cultivated plant species worldwide, such as maize (*Zea mays*), soybean (*Glycine max*), canola (*Brassica napus*), cotton (*Gossypium* spp.), alfalfa (*Medicago sativa*), potato (*Solanum tuberosum*), beetroot (*Beta vulgaris*), creeping bentgrass (*Agrostis stolonifera*), linseed (*Linum usitatissimum*), wheat (*Triticum aestivum*) and rice (*Oryza sativa*). The samples also included non-GMO crops used as negative controls. To assess the sensitivity of the assay, different concentrations of the same GM crops were used, typically ranging from 0.98 % to 100 % (ratio of the GM plant species in the total sample, expressed in mass/mass or copies/copies of haploid plant genomes). All our tested samples originated from certified reference materials (CRMs) provided under ISO17034

standards had been pre-validated for the presence of GM events and endogenous genes through qPCR testing. In addition to this certification, our laboratory also conducted extensive testing of these CRMs using official qPCR methods to confirm the presence of specified genes and ensure compliance with the reported standards. Detailed information about each sample and GMO events are provided in Table S1.

DNA extraction

DNA extractions were carried out by processing 200 mg of plant material with the NucleoSpin Plant II kit (Macherey-Nagel, GmbH, Germany) following the manufacturer's protocol. DNA concentrations of all samples were assessed using a NanoDrop One spectrophotometer and a Qubit (Thermo Scientific). Initial DNA concentration are reported in Table S1. We chose to standardise the DNA concentration of all our samples by diluting them in water to 50 ng µl−1. To explore the effects of low DNA input, we diluted one of the GMO maize samples in a dilution series starting from 50 ng µl−1 down to 5 ng µl−1 (10-fold), 0.5 ng µl−1 (100-fold) and 0.05 ng µl−1 (1000-fold). For downstream applications, we performed two additional replicates for 41 samples, including replicates of the dilution series (Table S1).

Recovery of target sequences and barcoding plant species

We retrieved target sequences for amplicon design from the EUginius (EUropean GMO INItiative for a Unified Database System) and the portugene (Moreira, Carneiro, and Pereira 2017b) GM sequence databases. Additionally, we added sequences encoding the dihydroflavonol 4-reductase gene from an unauthorized GM *Petunia* sequence (Fraiture et al. 2019). To avoid redundancy between the target sequences, we clustered nearly identical sequences and produced multiple alignments using Clustal Omega-v1.2.3 (Sievers and Higgins 2014). Aligned sequences were used to produce a consensus, retaining ambiguous bases and yielding a total of 115 unique sequences targeting specific GM events (Supplementary File S2). We also included primers for barcoding (*i.e.* species identification) purposes. For this, we added sets of primers for two widely used plant DNA barcoding loci (Kress 2017): the chloroplast-encoded

ribulose bisphosphate carboxylase large chain gene (rbcL**)** and the nuclear ribosomal internal transcribed spacer (ITS).

Amplicon design

For the 115 unique target sequences, we segmented sequences exceeding 300 bp into multiple regions for individual amplicon design to improve amplification yields and coverage of the GM events. After the segmentation, we obtained 230 candidate loci for primer design according to Fluidigm Inc. recommendations. The targeted amplicon length ranged from 62-240 bp reflecting constraints in conserved sequences and base composition. We obtained a total of 230 pairs of primers corresponding to the GM target sequences (Supplementary File S1). In the case of the barcodes that amplify rbcL and the ITS, we compiled a set of 27 primers representing various amplicon designs covering the same loci corresponding to rbcL and ITS (Supplementary File S3).

DNA sequencing library preparation

Libraries were prepared following the manufacturer's protocol PN 101-0414 G1 for the Juno LP 192.24 integrated fluidic circuits plate (IFC; Fluidigm Corporation, San Francisco, CA, United States). After loading all reagents on the IFC, target amplicons were generated for each sample through PCR amplification on a specialized thermocycler (Juno system; Fluidigm). A total of 257 primers were subdivided into 10X assay pools, with each pool containing primers for various targets, ensuring universal applicability of the assay independent of the samples analyzed. This pooling strategy, which was identical for every sample, was partially determined by Fluidigm Inc. as an optimal strategy given the specific oligos. Custom designed primers were added randomly to the pools. The IFC with an inlet containing 2 µl of sample pre-mix, 2 µl genomic DNA (50ng/µl) and 1 µl barcode primer mix consisting of a DNA sample and an individual barcode. High amplification efficiency of all 230 primer sets in the fluidic chip-based multiplex PCR was confirmed by the average sequencing depth of ~5000x, ensuring reliable and consistent target amplification across all samples. After amplification, samples were pooled in a single tube and purified. The first clean-up is double-sided (0.4X/0.9X Double-Sided SPRI), with first (0.4X) removing fragments that

are bigger than the targets, then (0.9X) binding and selecting targets by washing off the smaller fragments. The second and third clean-ups were to remove excess primers (0.8X SPRI). Finally, sequencing adapters were added by PCR to the purified library followed by a final round of purification according to the manufacturer's protocol. The quantity and quality of the library were assessed using a Qubit fluorometer assay (ThermoFisher) and a 4200 TapeStation electrophoresis instrument (Agilent). The final library was sequenced on a single lane of a NextSeq 500 system (Illumina) in mid-output mode adding ~30% PhiX to reduce issues due to low sequence complexity.

Recovering of plant barcode sequences and GM targeted sequences by mapping

We obtained the rbcL sequence of maize from NCBI (NC_001666.2) and the ITS sequence from potato (CP046695.1). Matching rbcL and ITS sequences were then retrieved using BLAST (Altschul et al. 1990) to complete a library of barcoding sequences for all eleven included plant species. The crop barcode sequences were added to the GM reference sequences. We demultiplexed raw read data using bcl2fastq v-2.19.0.316 and used trimmomatic v-0.36 (Bolger, Lohse, and Usadel 2014) for quality trimming. The trimming parameters used were as follows: adapter clipping with ILLUMINACLIP:TruSeq3-PE.fa:2:30:10, removal of leading low-quality bases with LEADING:15, removal of trailing low-quality bases with TRAILING:15, a sliding window trimming approach to cut when the average quality within a window of 5 bases falls below 15 (SLIDINGWINDOW:5:15), and dropping reads shorter than 50 bases (MINLEN:50). Forward and reverse reads were merged using flash v-1.2.11 (Magoč and Salzberg 2011). We aligned merged reads to the reference sequences using bowtie2 v-2.3.5 using the following settings: --very-sensitive-local --phred33 (Langmead et al. 2019). Based on the aligned reads, we calculated the depth per position ignoring locations outside of designed amplicons. We accessed all nucleotide sequences for rbcL and ITS available on NCBI for each of the eleven plant species generating eleven fasta files for rbcL and eleven fasta files for ITS. The fasta files were used as a reference to align the merged reads by bowtie2 v-2.3.5 using the following settings: --very-sensitive-local --phred33 (Langmead et al. 2019). From the aligned reads, we calculated the depth per position using samtools v-1.19. and we compared the depth to

the previous depth of the rbcL sequence from maize (NC_001666.2) and the ITS sequence from potato (CP046695.1) with the retrieved BLAST sequences of the eleven crops. NC_001666.2 and CP046695.1 reference sequences were compared based on mapped reads per sample. We document our analyses pipeline in Supplementary File S4. All statistical analyses were performed with the R statistical software version 4.3.3.

RESULTS

Sequence-guided design of the amplicon sequencing assay

In order to test the feasibility and assess the performances of the amplicon-sequencing assay to detect GM material in food, feed or seed matrices, we gathered a total of 115 consensus GM sequences from various sources: EUginius (EUropean GMO INItiative for a Unified Database System), and portugene (Moreira, Carneiro, and Pereira 2017b) databases (Figure 1A). In addition, we manually added three sequences from a gene encoding the dihydroflavonol 4-reductase from an unauthorized GM *Petunia* recently detected in the (Fraiture et al. 2019). For the design of amplicons, we retrieved the GM consensus sequences if multiple redundant sequences were found in databases. The lengths of these sequences ranged from 79 bp to 24,595 bp, with an average length of 1856 bp. Out of these, 99 sequences were longer than 300 bp. For sequences longer than 300 bp, multiple amplicons were designed to cover the entire target region. Specifically, the sequences longer than 300 bp were fragmented into smaller segments to ensure efficient amplification and coverage of the GM events, hence we designed multiple similarly spaced amplicons to cover the entire event (Figure 1C). Based on this GM sequence set, 230 primer pairs for GM sequences that correspond to 82 unique GM sequences were designed (Figure 1C) and 15 rbcL + 12 ITS primers were added for plant species identification (Figure 1B). Following the manufacturer's instructions, we aimed for an amplicon length of ~200bp and we obtained a maximum length of 240 bp for 14 GM locus and the smallest amplicons with a length in the range of 69-199 bp for 21 GM loci (Figure 2A). All 230 primer sets demonstrated high amplification efficiency in

fluidic chip-based multiplex PCR, with an average sequencing depth of ~5000x consistent with robust amplification performance.

The amplicon sequencing was performed using a microfluidics platform (Figure 1C). In a single flowcell run, we analysed a total of 186 samples including replicates, representing 92 distinct samples of GMO and non-GMO control material. The species covered were alfalfa (n=10), beetroot (n=6), canola (n=23), cotton (n=26), creeping beetroot (n=3), linseed (n=3), maize (n=66), potato (n=6), rice (n=3), soybean (n=37) and wheat (n=3) (Figure 2B).

Illumina NextSeq 550 sequencing generated a total of 80,478,507 reads after trimming and merging read pairs per sample and locus. Furthermore, for every sample, the reads were aligned to the 115 target sequences and plant barcoding sequences rbcL and ITS sequences of the 11 species. One sample of maize obtained a maximum of 3,122,899 aligned reads (Figure 2C). The minimum number was 229 aligned reads which corresponds to a sample of soybean (Figure 2C). The other three minima of aligned reads ranged from 663 to 1,418 and correspond to 3 replicates of DNA dilution (1000x) of a maize sample (Figure 2B).

Identification of plant species by amplification of plant barcoding loci

An important step for the identification of GMO crops is to determine species identities. We incorporated in our targeting sequencing assay two sets of primers (Kress 2017) previously recognized as plant DNA barcodes that amplify the rbcL and ITS loci. Given that our 15 rbcL + 12 ITS primers are amplifying in different regions of rbcL and ITS (Figure 3A), we decided to work with the maximum number of mapped reads per barcode, which were correlated to the mean (r = 0.897***, Pearson correlation coefficient), median (r = 0.848***) and mode (r = 0.848***). To evaluate the success of rbcL and ITS amplification, we extracted the maximum number of mapped reads for the 186 samples. The rbcL maximum count was significantly positively correlated to ITS maximum counts for alfalfa, canola, cotton, potato and soybean in the range of 0.983*** (p-value <0.001) to 0.748*** (Figure 3B). In the cases of beetroot, creeping beetroot,

rice and wheat, the correlation was not significant (*p*-values > 0.05), given that we have a low number of samples (*n* = 3-6; Figure 3B). Maize samples showed also a positive correlation with 0.503*** (Figure 3B). We also compared the maximum read count at barcodes against the total aligned counts to ensure that the general quality of the samples has been captured by the plant barcodes. For ITS, 10 plant species showed a significant and positive correlation from 0.88*** to 1***, with the only exception of beetroot lacking a strong correlation (0.25*). RbcL varies more between species with a range of -0.96 p=0.19 to 0.97***.

Every sample was mapped to the 11 crop species sequences of rbcL and ITS. We assigned species identities according to the highest read depth among rbcL and ITS reference sequences. As the species was known for all samples, we assessed the power of the assay to recover the true species. Only using, the ITS primers we achieved 91 out of 186 correct predictions of the species identity corresponding to an accuracy of 48.9%. For the rbcL primers, we achieved 182 out of 186 correct predictions of the species identity corresponding to an accuracy of 97.8%. Hence, rbcL primers predicted more accurately the species and we used this barcoding locus for most species (Figure 3C). The species that performed best was maize showing the highest correctly mapped reads ratio for 13 samples (79-172), meaning that the primers used to amplify rbcL and the rbcL sequence used to align are providing the best correct mapping. The worst correctly mapped reads ratios correspond to the three rice samples: 0.912- 0.686.

Detection of GM events based on amplicon sequencing

The GM positive controls utilized in this study were sourced from CRM that have already been rigorously tested for endogenous reference gene amplification by the national certifying bodies. This pre-validation ensures the presence and accuracy of GM events and eliminating the need for repetitive endogenous gene testing in our experimental framework. Notably, each CRM used in this study has been verified for the accuracy of GM content by our laboratory using official qPCR methods. We analyzed 98 individual samples including 11 non-GMO controls. The samples were known to carry at least 10 sequence fragments from a specific GM event with 9 samples originating from

maize and one from cotton (Supplementary Table S2). We analyzed the 10 GMO samples and compared these against the non-GMO controls. To account for variation in total reads among samples and the uneven presence of GM sequences, we normalized read counts using normalized mean depth by the maximum read depth at the barcoding locus. In the case of maize, we used rbcL because the read depth was higher in comparison to ITS (Figure 3C). For cotton, we used ITS because the read depth was higher for ITS (Figure 3C). From the 10 analyzed GM events, six events (MON89034, BT11, DAS59122, MIR162, MON88017 and MON87427) amplified well in the GMO sample and showed no meaningful amplification in the non-GMO control (Figure 4A-F). The high amplification efficiency of all 230 primer sets in the fluidic chip-based multiplex PCR was evidenced by the consistent and robust detection of GM sequences by the average sequencing depth of ~5000x. For example, MON 89034 amplified well for 3'MON89034 (normalized count 0.094), 5'MON89034 (normalized count 0.004), hsp70-locus56 (normalized count 0.747), hsp70-locus307 (normalized count 0.285), LTa.lhcbl (normalized count 0.002), tahsp17 (normalized count 0.331) and CTP2 (normalized count 0.405). For the three other GM events (MON810, SYN3272 and MON87460), the GMO maize also amplified more GM sequences compared to the non-GMO control (Figures 4G-I). For example, the sample MON810 where the non-GMO control is amplified in 5' MON810-locus 59 (normalized count 0.383) vs the GM sample (normalized count: 0.526) (Figure 4G). The last sample corresponds to MON15985 where the sequence labelled "Cotton_MON15985" amplified more in the GMO cotton (0.148) compared to non-GMO cotton (0.046 normalized counts; Figure 4J). The most discriminant amplicon for detecting the MON15985 event was the Oriv amplicon.

Specificity and sensitivity of GMO detection

For monitoring purposes, mixed GMO samples need to be screened. To assess the power of our targeted sequencing assay, we tested three GMO samples diluted in control DNA at various ratios. For the GMO SYN3272 at the locus amy797E, the GMO (constituting 9.8%) amplified in the range of 0.01- 0.002 normalized read counts contrasted with no recovered reads for the GMO mixed in at 0.98% (Figure 5B). The

sample DAS59122 amplified in both the 1% mixture in the range of 1.15 e-10 – 0.0021 and the 10% mixture in the range of 0.00018 – 0.00796 (Figure 5C). Unexpectedly, MON810 showed more amplification for GMO 1% compared to GMO 10% (Figure 5A). For example, 5' MON810 locus 336 amplified for GMO 10% 1.2830 compared to 1.553 normalized read counts amplified for GMO 1%. The non-GMO control sample showed however 0.7572 normalized read counts. Next, we assessed the sensitivity of the amplicon assay to detect DNA diluted 10x, 100x, and 1000x. We tested the sample MON88017, however, the normalized counts mismatched the expected trend from the dilutions. The most likely explanation for this variability is the overall low number of reads likely inducing noise in read counts (Figure 5D). Using the mean depth of mapped reads (before normalization), we recovered the expected change in read depth along the dilution series (Figure 5E).

Additional amplicons for exploratory detection of GM events

We recovered 73 target sequences from EUginius, portugene and Fraiture *et al.* (2019) and amplified these in using the custom amplicon sequencing assay. We assessed amplification of the 10 GM sequences described above (Figure 6). MON15985 cotton is amplifying the sequence Cotton_MON15985 as expected, and in addition OriV (Figure 4J). Non-GMO controls showed amplification of multiple 3' and 5' GM flanking sequences as expected from the above findings. Three additional sequences including "FB707511.1 cry1A.105", "DL476427.1 CORN EVENT" and "aadA" showed amplification in the MON15985 GMO. Regarding the GMO maize samples, MON 89034 amplified the sequences 3'MON89034, hsp70, tahsp and CTP2 (Figure 4A). Beyond amplified flanking sequences, three sequences amplified in the GMO including "FB707511.1 cry1A.105" and "DL476427.1 CORN EVENT", and "dihydroflavonol4-reductase_MF521566.1", which the MON15985 GMO was not known to contain.

DISCUSSION

Quantative PCR has been for long the widely accepted standard for GMO detection in all matrices that could be circulating as food, feed, seeds and in the environment. qPCR is highly sensitivity but labour-intensive. Improving methods for detecting GMOs is therefore important. Here, we assessed the feasibility of a microfluidics-based approach for the detection of GM events. We desigend a set of 230 amplicons that represent 82 unique GM events. In addition, 27 plant barcoding primers allowed to determine species contained in the samples. The advantage of using amplicon sequencing is the versatility and expandability to perform monitoring of samples without prior knowledge of the genetic modifications present.

Primers to amplify the barcoding genes rbcL and ITS allowed to identify most plant species contained in the sample. Samples originating from cotton were however not well identified with low read counts mapping to the rbcL gene. Low coverage may be due to selected primers performing poorly in those samples. The complexity of amplifying a barcoding locus using multiple overlapping pairs of primers was apparent in the read mapping patterns along the rbcL and ITS sequences. Replacing the pool of barcoding loci primers with pairs of custom-designed primers for a specific range of species would alleviate the complexity in amplification and likely increase consistency in amplification across the desired species. The detection of GM events in positive control samples was successful for a wide range of GM sequences. However, non-GMO controls amplified in some GM regions including 5’ and 3' flanking sequences, the promoters, and terminators, which likely have homology to regions in the genome outside of the GM sequences. Such a lack of specificity in flanking and promoter sequences of GM events against regular plant DNA sequences can explain why GMO event promoters are also amplifying in non-GMO crops (Jores et al. 2021). The GM regions, which were amplified in non-GMO samples showed indeed sequence homology in genomes of non-GMO crops. This supports the lack of specificity in the 5’ and 3' flanking sequences of GM events. However, the recovered homologous sequences in crop genomes did not present 100% identity, which suggests that SNP calling could be used to differentiate between reads mapping to a GM event versus reads mapping to non-GMO sequences elsewhere in the genome.

Endogenous reference genes are conventionally used as controls in GM crop detection to ensure the presence of target DNA. Here, we opted for a different approach that extends beyond species-specific restrictions of amplifiable loci (Huang et al. 2013). Given the diverse nature of the plant materials to be tested, using typical endogenous reference genes would have constrained our ability to detect GM sequences in diverse monocot and dicot crops. Relying on barcoding loci such as chloroplast or ribosomal DNA markers makes amplification possible in principle across entire kingdoms. However, primer specificity and amplification biases need to be carefully evaluated. In our assessment, amplification of barcoding loci provided reasonable quantification accuracy. This is supported by the fact that our study relied on CRM, which had already undergone qPCR testing for endogenous gene presence. Our microfluidic approach could well identify species identities. However, we advocate for caution if even more diverse plant species are to be assayed. A preliminary trial run to compare amplification efficiency across species and taxonomic assignments is recommended. The microfluidic amplicons developed in this study diverge also in other ways from traditional GM detection approaches. Notably, we avoided host-exogenous DNA boundaries in our primer design. Such cross-boundary qPCR is typically employed to confirm the presence of a specific insertion event in a specific genetic background. However, designing amplicons only on exogenous DNA allows for universal amplification of specific GM sequences at any insertion locus or any genetic background. For specific host-exogenous DNA boundary detection our assay could either be expanded by additional amplicons or supplemented by qPCR validation.

Previous studies describing multiplex detection of GM events were largely based on amplicons stemming from known primer pairs used in routine GMO food and feed sample screening strategies and approved in GMO reference material (Scholtens et al. 2017, Arulandhu et al. 2018). In some jurisdictions including the European Union, there is an obligation to proivde a detection method specific to the authorized GM event. As the common practice usually is centred on the use of qPCR, often only primer sequences are made available, limiting the investigation of GM events. The dependency

on a certified set of primers restricts the completeness of the GM event sequences to be recovered, because amplicon sequences do not cover the complete GM event. The lack of long amplicon sequences and reference material accessibility are challenging to investigate GMO sequences comprehensively (Moreira, Carneiro, and Pereira 2017b). Our study shows that a *de novo* design of GM amplicons is feasible using public sequence information and that a broad range of potential GM sequences can be assessed in parallel. The approach of a microfluidics-based targeted amplicon sequencing assay enables to screen both hundreds of samples (or replicates) simultaneously but also allows for large sets of primers to be included in parallel. In principle, the microfluidics chips would allow the pooling of thousands of primer pairs for single-step amplifications. Given the uncertainty of amplifying specific sequences from unknown samples and modification events, the sequence information provides significantly greater certainty about the identity of an amplified sequence. This contrasts with qPCR approaches that lack direct validation capabilities under non-standard conditions. In contrast, targeted amplicon sequencing is less sensitive compared to qPCR. In our analyses, we found reliable amplification up to ~1:100 dilutions, at lower concentrations the detection is likely to be only poorly reproducible. Such detection limitations can be remedied partially by increasing the overall sequencing coverage of the amplicons as sensitivity is at least partially correlated with sequencing depth.

This study demonstrates the applicaton of a comprehensive amplicon sequencing assay that leverages the parallelization offered by microfluidics platforms and the depth of NGS for the detection of GMOs. Our work fits into efforts to standardize and propose a statistical framework for the detection of GMOs based on the number of reads aligned per sample. Our workflow expands the capabilities by targeting many sequences of interest specifically and allowing for the efficient detection of plant species present in a sample. With 230 designed amplicons corresponding to GM events and additional primers for species-specific barcoding, our approach represents a significant advancement in GMO detection. The assay robustness was demonstrated by successfully identifying ten known genetic modifications across different crop species and showcasing the potential for uncovering undocumented genetic events. The

integration of rbcL and ITS barcoding primers enhanced the assay's precision by enabling accurate species identification within mixed samples, an essential step in standardised GMO presence across samples. Challenges remain in distinguishing between GM events and homologous non-GMO sequences. The use of sequence information provided by NGS offers a direct validation of the detected genetic material. The successful identification of GMOs in this study underscores the importance of developing advanced screening methods. Our study shows the relevance of amplicon sequencing that can be realistically implemented into GMO detection and efficiently analyzed using a structured bioinformatics pipeline.

Data availability

Raw sequencing data is available from the NCBI Sequence Read Archive in BioProject PRJNA1117073.

Competing interests

CSRA and DC have filed for a European patent related to the microfluidics-based detection of genetically modified organisms (GMOs).

Funding

DC received funding by the Swiss Federal Offices for Agriculture and for the Environment (FOEN).

Acknowledgments

Data was generated in collaboration with the Genetic Diversity Centre (GDC), ETH Zurich.

Figures

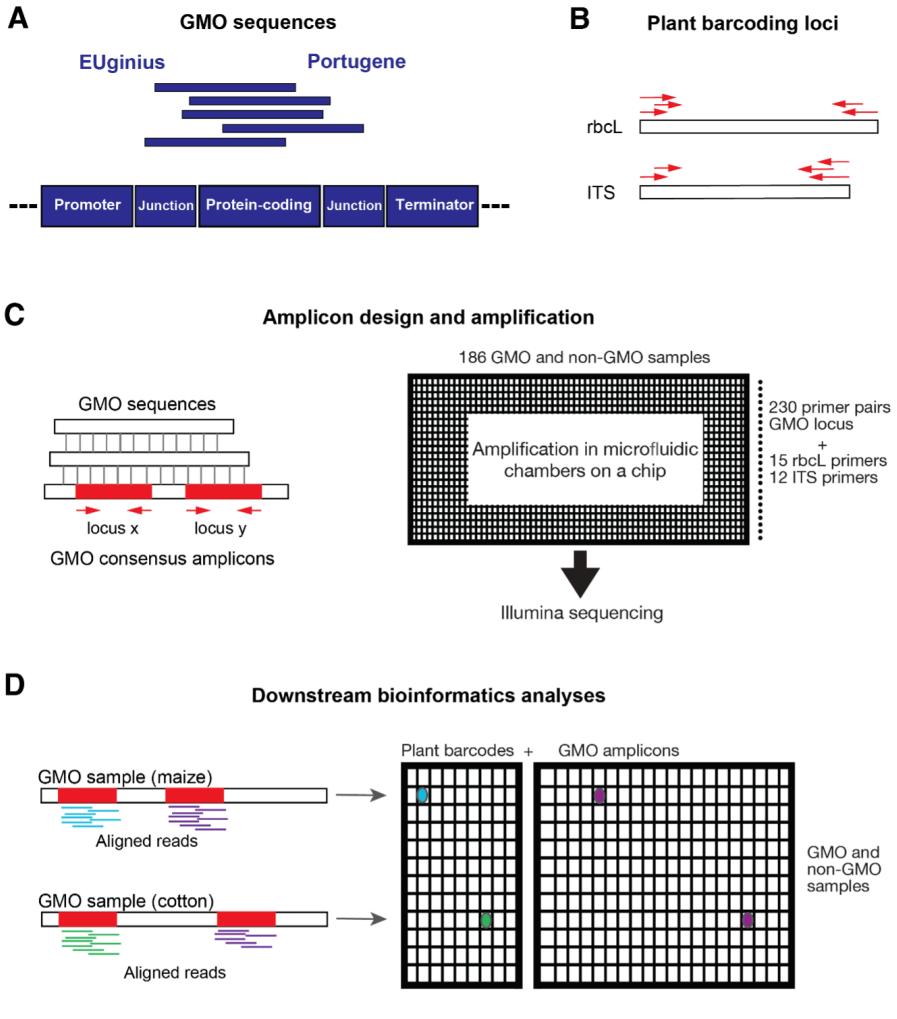

Figure 1. Design and workflow of the targeted amplicon sequencing loci. (A) Recovery of GM sequence events from EUginius and Portugene databases, and the general structure of a GM event, consisting typically of promotor, junctions, protein-coding and terminator elements. (B) Multi-primer design for plant barcoding loci to cover sequence diversity among crop plants. (C) Multiple alignment of GMO sequences per event to obtain consensus sequences per locus in order to design consensus amplicons of around ~300bp. The microfluidics-based Juno system was used to amplify all amplicons across all samples in parallel. (D) Bioinformatics downstream analyses including alignment of reads against plant barcoding and GMO consensus sequences.

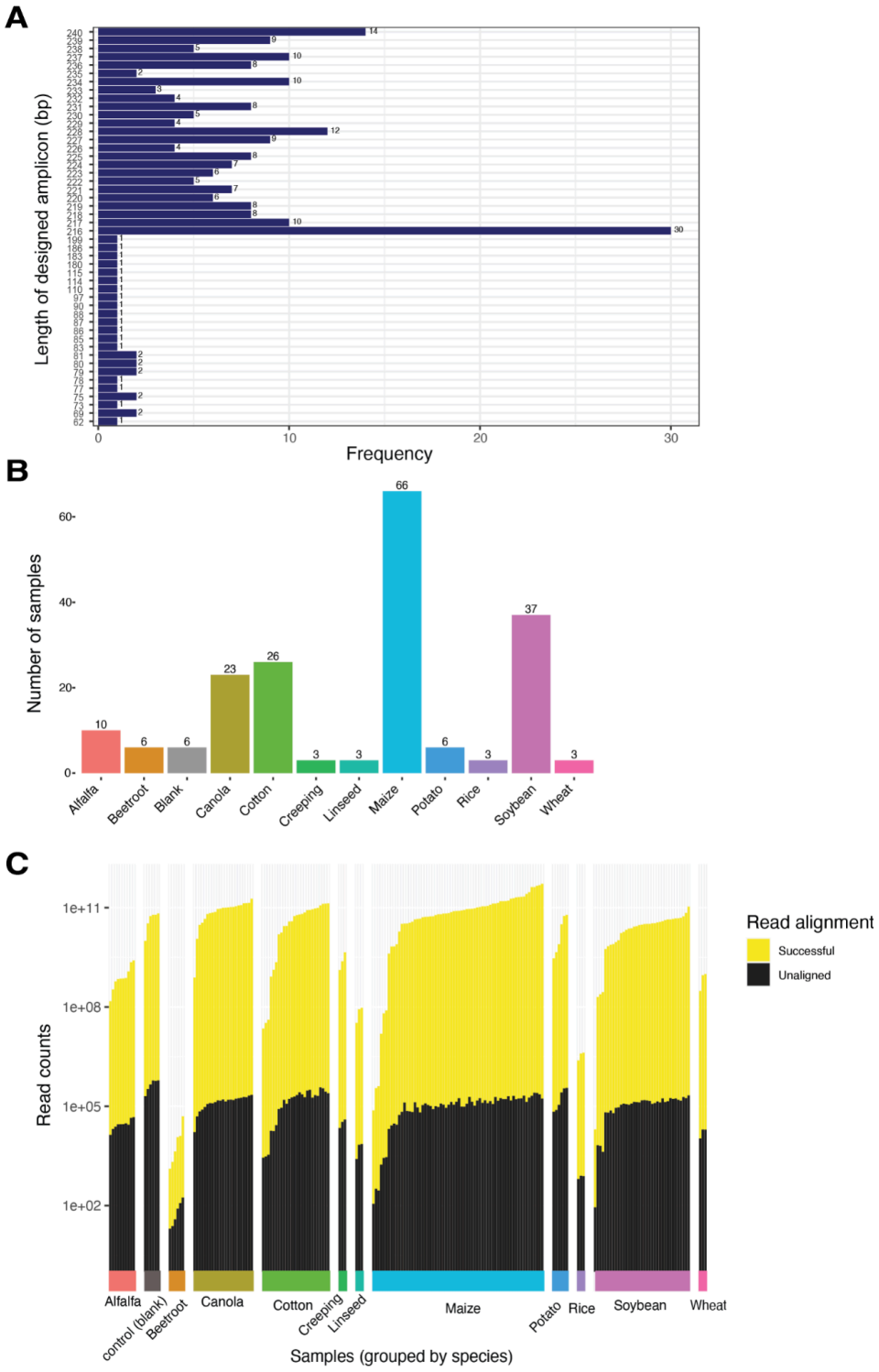

Figure 2. Overview of the designed amplicons, samples used for testing and amplification. (A) Sequence length distribution among the 230 designed GM amplicons. (B) Number of samples per plant species covering 92 different GMO events across multiple samples. (c) Mapped and unmapped reads recovered per individual sample and species.

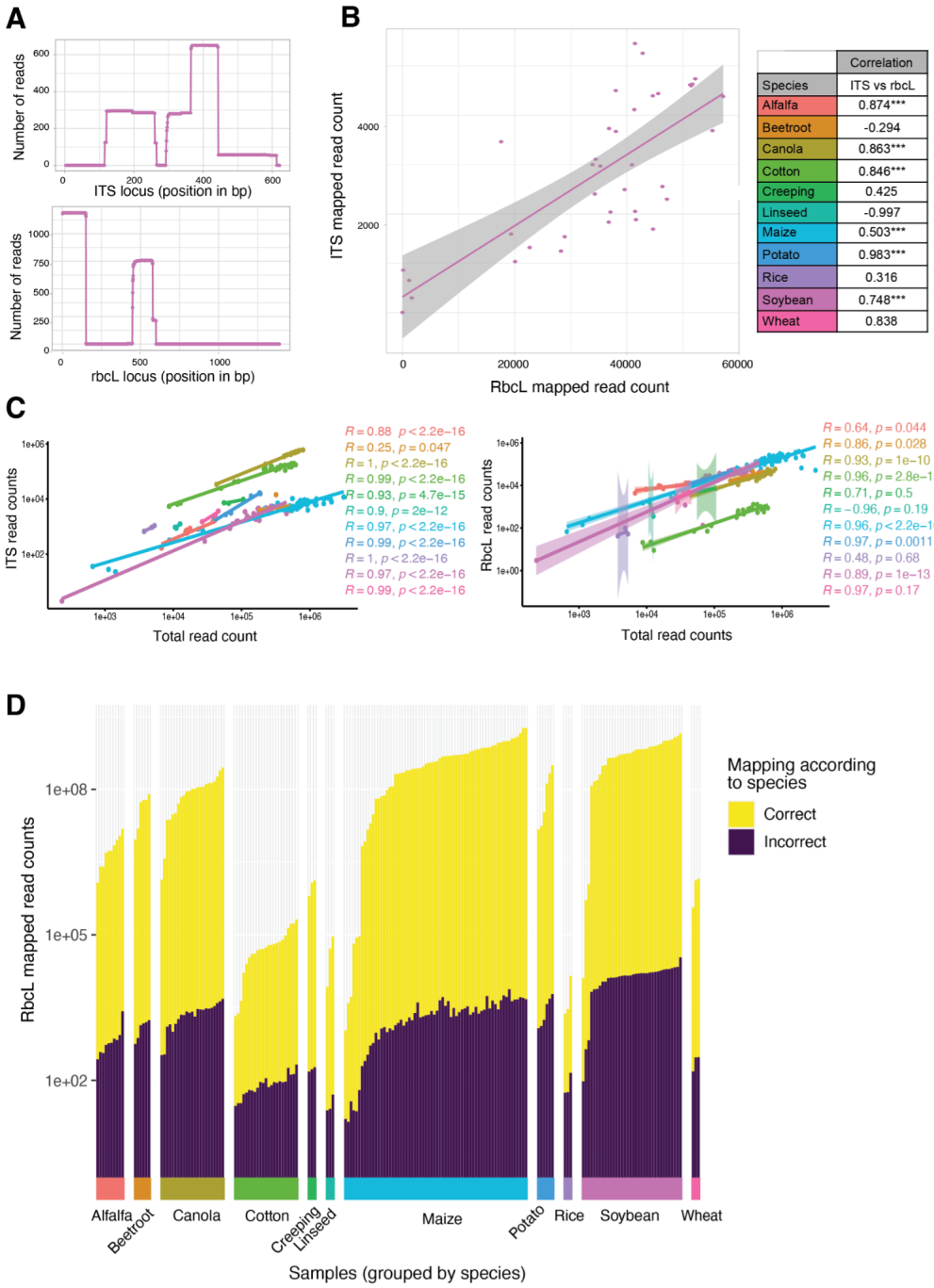

Figure 3. Assessment of read mapping distribution on the plant barcoding reference sequences for the ITS and rbcL loci. (A) Depth of mapped reads at the ITS and rbcL loci for a soybean sample. (B) Correlations of the maximum mapped read count for ITS compared to rbcL for all soybean samples including a table of correlations for other plant species. (C) Correlation of the mapped read counts per plant barcoding locus compared to the total read counts. Correlation of total aligned read counts compared against the maximum count of rbcL reads showing *r* and *p*-values. (D) Counts of mapped reads for the rbcL locus mapped to the correct species reference sequence versus other reference sequences (note the logarithmic scale).

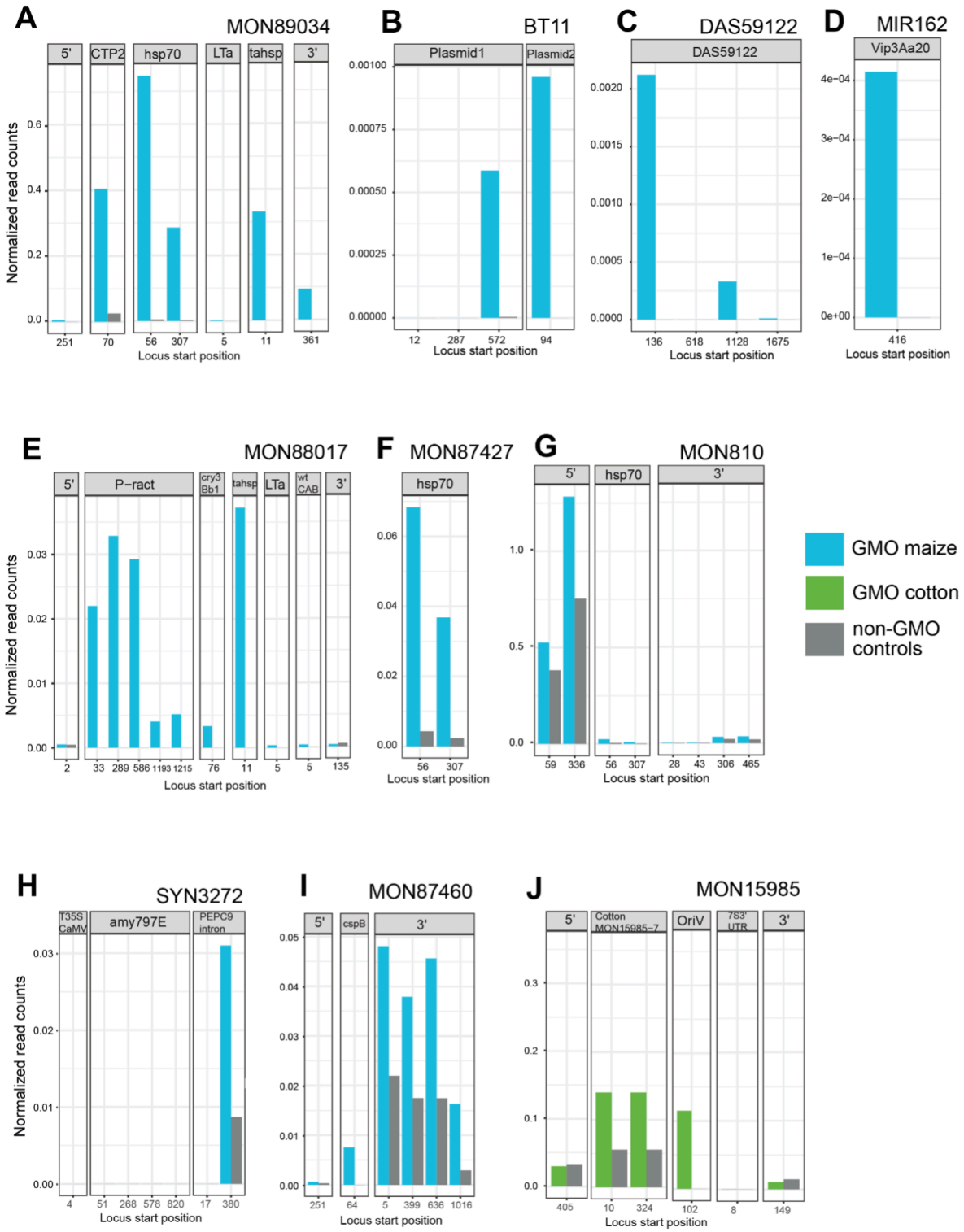

Figure 4. Assessment of known GMO events by amplicon sequencing and contrast to non-GMO control samples. The following GM events were covered with multiple amplicons. Locus start positions indicate the amplicon location on the GMO consensus sequence. Events include (A) MON89034, (B) BT11, (C) DAS59122, (D) MON88017, (E) MON88017, (F) MON87427, (G) MON810, (H) SYN3272, (I) MON87460 and (J) MON15985.

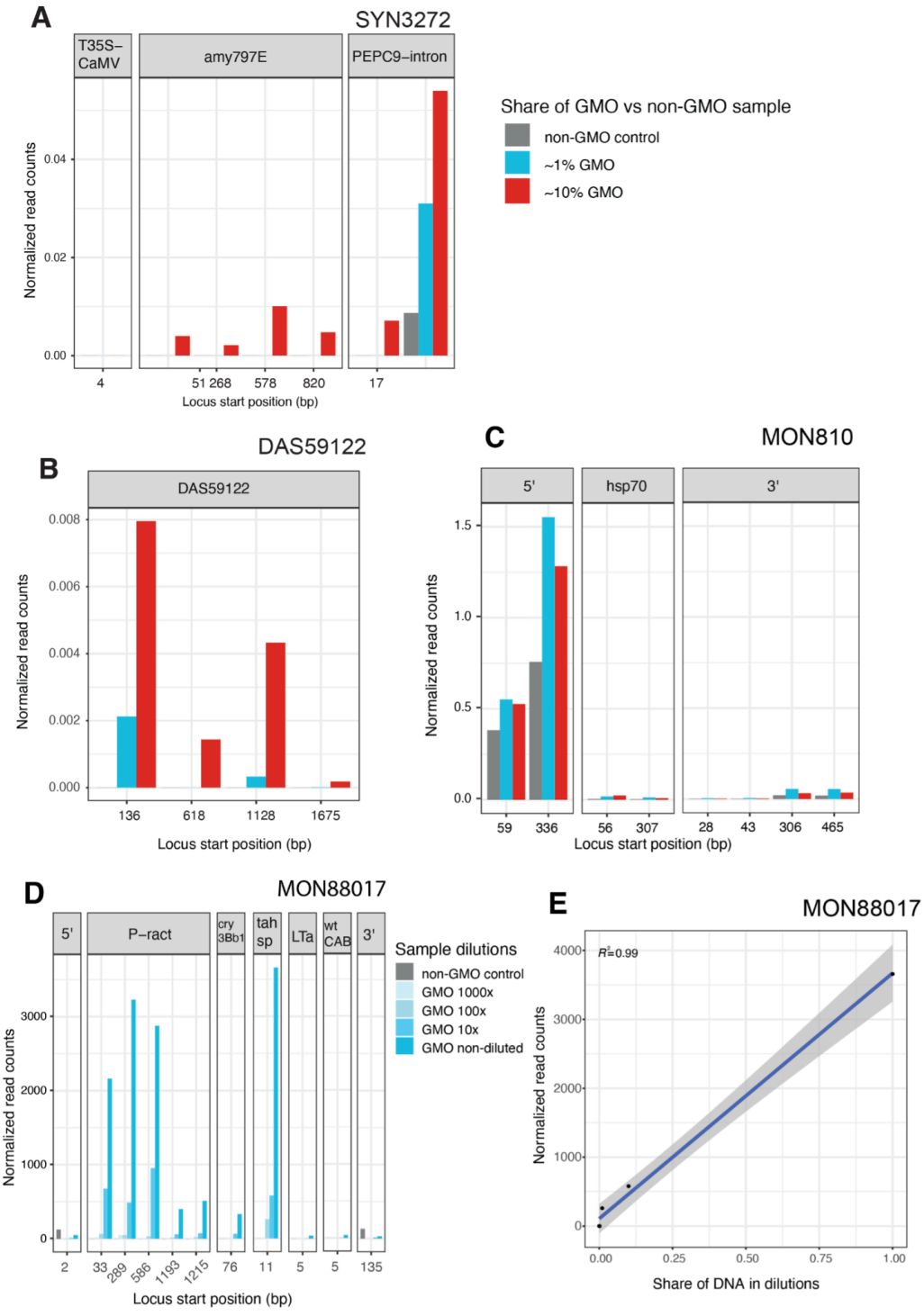

Figure 5. Assessment of DNA concentration and sample dilution effects. (A) GM event called SYN3272, composed of three recovered regions (amy797E, PEPC9-intron, and T35S-CaMV) across different positions of the GM locus. Read counts were adjusted for each of the samles diluted to contain 0.98% concentration of a GMO sample in blue, the 9.8% concentration of GMO maize in red and the non-GMO maize in grey. (B) GM event DAS59122 was amplified at four different loci. (C) GM event MON810 spanning three assessed regions (3', 5' and hsp70). (D) Assessment of sample dilution effects (GMO samples mixed with sterile water) for GM event MON88017. (E) Correlation of GMO DNA share in diluted sampled with the normalized read counts.

Figure 6. Amplifications of GM events included in the databases EUginius, Portugene and provided by Fraiture *et al.* (2019). Heatmaps for 11 samples covering nine different GMO events for maize with one non-GMO maize as control; one GMO event for cotton with one non-GMO cotton as control, aligned to the 82 reference sequences generated from the three databases (EUginius, portugene and Fraiture) and read counts were normalized using barcoding loci sequencing depth. The red stars are marking amplification in the samples for MON15985 GMO cotton where "FB707511.1 cry1A.105", "DL476427.1 CORN EVENT" and "aadA" were amplified; for the MON 89034 "FB707511.1 cry1A.105" and "DL476427.1 CORN EVENT" and "dihydroflavonol4-reductase_MF521566.1".

References

Alasaad, Noor, Hussein Alzubi, and Ahmad Abdul Kader. 2016. “Data in Support of the Detection of Genetically Modified Organisms (GMOs) in Food and Feed Samples.” *Data in Brief* 7 (June):243–52. https://doi.org/10.1016/j.dib.2016.02.035.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Arulandhu, Alfred J., Jeroen van Dijk, Martijn Staats, Rico Hagelaar, Marleen Voorhuijzen, Bonnie Molenaar, Richard van Hoof, et al. 2018. “NGS-Based Amplicon Sequencing Approach; towards a New Era in GMO Screening and Detection.” *Food Control* 93 (November):201–10. https://doi.org/10.1016/j.foodcont.2018.06.014.

Aubry, Sylvain, Sarai Reyes Avila, Daniel Croll, and Bastien Christ. 2021. “Chapter 10. Omics-Based Detection, Identification and Quantification of GM Food and Feed: Current Challenges and Perspectives.” In *Food Chemistry, Function and Analysis*, edited by Jorge Barros-Velázquez, 257–70. Cambridge: Royal Society of Chemistry. https://doi.org/10.1039/9781839163005-00257.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics (Oxford, England)* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Dong, Cheng, Fei Li, Yun Sun, Dongling Long, Chunzhao Chen, Mengyan Li, Tao Wei, Rui P. Martins, Tianlan Chen, and Pui-In Mak. 2023. “A Syndromic Diagnostic Assay on a Macrochannel-to-Digital Microfluidic Platform for Automatic Identification of Multiple Respiratory Pathogens.” *Lab on a Chip*, November. https://doi.org/10.1039/D3LC00728F.

Dong, Wei, Litao Yang, Kailin Shen, Banghyun Kim, Gijs A. Kleter, Hans JP Marvin, Rong Guo, Wanqi Liang, and Dabing Zhang. 2008. “GMDD: A Database of GMO Detection Methods.” *BMC Bioinformatics* 9 (1): 260. https://doi.org/10.1186/1471-2105-9-260.

Endres, A Bryan. n.d. “‘GMO:’ Genetically Modified Organism or Gigantic Monetary Obligation? The Liability Schemes for GMO Damage in the United States and the European Union,” 55.

Fraiture, Marie-Alice, Gabriella Ujhelyi, Jaroslava Ovesná, Dirk Van Geel, Sigrid De Keersmaecker, Assia Saltykova, Nina Papazova, and Nancy H. C. Roosens. 2019. “MinION Sequencing Technology to Characterize Unauthorized GM Petunia Plants

Circulating on the European Union Market." *Scientific Reports* 9 (1): 7141. https://doi.org/10.1038/s41598-019-43463-5.

"GMO Content - By Analyte Group - Certified Reference Materials Catalogue of the JRC." n.d. Accessed May 28, 2024. https://crm.jrc.ec.europa.eu/c/By-analyte-group/GMO-content/40481/.

Ho, Kuan-Lun, Jing Ding, Jia-Shao Fan, Wai Ning Tiffany Tsui, Jianfa Bai, and Shih-Kang Fan. 2023. "Digital Microfluidic Multiplex RT-qPCR for SARS-CoV-2 Detection and Variants Discrimination." *Micromachines* 14 (8): 1627. https://doi.org/10.3390/mi14081627.

Huang, Huali, Fang Cheng, Ruoan Wang, Dabing Zhang, and Litao Yang. 2013. "Evaluation of Four Endogenous Reference Genes and Their Real-Time PCR Assays for Common Wheat Quantification in GMOs Detection." *PLOS ONE* 8 (9): e75850. https://doi.org/10.1371/journal.pone.0075850.

Jagadeesan, Balamurugan, Peter Gerner-Smidt, Marc W. Allard, Sébastien Leuillet, Anett Winkler, Yinghua Xiao, Samuel Chaffron, et al. 2019. "The Use of next Generation Sequencing for Improving Food Safety: Translation into Practice." *Food Microbiology* 79 (June):96–115. https://doi.org/10.1016/j.fm.2018.11.005.

Jores, Tobias, Jackson Tonnies, Travis Wrightsman, Edward S. Buckler, Josh T. Cuperus, Stanley Fields, and Christine Queitsch. 2021. "Synthetic Promoter Designs Enabled by a Comprehensive Analysis of Plant Core Promoters." *Nature Plants* 7 (6): 842–55. https://doi.org/10.1038/s41477-021-00932-y.

Kress, W. John. 2017. "Plant DNA Barcodes: Applications Today and in the Future." *Journal of Systematics and Evolution* 55 (4): 291–307. https://doi.org/10.1111/jse.12254.

Langmead, Ben, Christopher Wilks, Valentin Antonescu, and Rone Charles. 2019. "Scaling Read Aligners to Hundreds of Threads on General-Purpose Processors." Edited by John Hancock. *Bioinformatics* 35 (3): 421–32. https://doi.org/10.1093/bioinformatics/bty648.

Lehnert, Thomas, and Martin A. M. Gijs. 2024. "Microfluidic Systems for Infectious Disease Diagnostics." *Lab on a Chip* 24 (5): 1441–93. https://doi.org/10.1039/D4LC00117F.

Magoč, Tanja, and Steven L. Salzberg. 2011. "FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies." *Bioinformatics (Oxford, England)* 27 (21): 2957–63. https://doi.org/10.1093/bioinformatics/btr507.

Moreira, Filipa, João Carneiro, and Filipe Pereira. 2017a. "A Proposal for Standardization of Transgenic Reference Sequences Used in Food Forensics." *Forensic Science International: Genetics* 29 (July):e26–28. https://doi.org/10.1016/j.fsigen.2017.04.022.

———. 2017b. “A Proposal for Standardization of Transgenic Reference Sequences Used in Food Forensics.” *Forensic Science International: Genetics* 29 (July):e26–28. https://doi.org/10.1016/j.fsigen.2017.04.022.

Nouwairi, Renna L., Larissa L. Cunha, Rachelle Turiello, Orion Scott, Jeff Hickey, Scott Thomson, Stuart Knowles, Jeff D. Chapman, and James P. Landers. 2022. “Ultra-Rapid Real-Time Microfluidic RT-PCR Instrument for Nucleic Acid Analysis.” *Lab on a Chip* 22 (18): 3424–35. https://doi.org/10.1039/D2LC00495J.

Scholtens, Ingrid M. J., Bonnie Molenaar, Richard A. van Hoof, Stephanie Zaaijer, Theo W. Prins, and Esther J. Kok. 2017. “Semiautomated TaqMan PCR Screening of GMO Labelled Samples for (Unauthorised) GMOs.” *Analytical and Bioanalytical Chemistry* 409 (15): 3877–89. https://doi.org/10.1007/s00216-017-0333-7.

Serageldin, I. 1999. “Biotechnology and Food Security in the 21st Century.” *Science* 285 (5426): 387–89. https://doi.org/10.1126/science.285.5426.387.

Sievers, Fabian, and Desmond G. Higgins. 2014. “Clustal Omega.” *Current Protocols in Bioinformatics* 48 (December):3.13.1-3.13.16. https://doi.org/10.1002/0471250953.bi0313s48.

Willems, Sander, Marie-Alice Fraiture, Dieter Deforce, Sigrid C. J. De Keersmaecker, Marc De Loose, Tom Ruttink, Philippe Herman, Filip Van Nieuwerburgh, and Nancy Roosens. 2016. “Statistical Framework for Detection of Genetically Modified Organisms Based on Next Generation Sequencing.” *Food Chemistry* 192 (February):788–98. https://doi.org/10.1016/j.foodchem.2015.07.074.

Xue, Bantong, Jinlong Guo, Youxiong Que, Zhiwei Fu, Luguang Wu, and Liping Xu. 2014. “Selection of Suitable Endogenous Reference Genes for Relative Copy Number Detection in Sugarcane.” *International Journal of Molecular Sciences* 15 (5): 8846–62. https://doi.org/10.3390/ijms15058846.

Yang, Bo, Ping Wang, Zhenqing Li, Qingxiang You, Shinichi Sekine, Junshan Ma, Songlin Zhuang, Dawei Zhang, and Yoshinori Yamaguchi. 2023. “Simultaneous Amplification of DNA in a Multiplex Circular Array Shaped Continuous Flow PCR Microfluidic Chip for On-Site Detection of Bacterial.” *Lab on a Chip* 23 (11): 2633–39. https://doi.org/10.1039/D3LC00274H.

CHAPTER 2

LOCO: LOw depth COpy algorithm to infer the ancestry of an admixed individual without reference panels

C. Sarai Reyes-Avila [1], Madleina Caduff [2], Daniel Croll[1], Daniel Wegmann[2]

[1]Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, CH-2000, Neuchâtel, Switzerland

[2] Statistical and Computational Biology, Department of Biology, University of Fribourg, CH-1700, Fribourg, Switzerland

ABSTRACT: Detecting genetic ancestry in admixed individuals is important for understanding population history. Predefined reference panels from well-characterized ancestral populations are relevant for current ancestry inference methods. However, these reference panels are frequently missing or incomplete, especially in ancient samples or poorly represented organisms. In this study, we present LOCO (LOw depth COpy algorithm), a novel probabilistic framework to infer the ancestry of admixed individuals directly from genotype likelihoods derived from low-coverage or low-quality sequencing. Our method builds upon Li & Stephens copy models (N. Li & Stephens, 2003) as implemented in HAPMIX (Price et al., 2009) and RASPberry (Wegmann et al., 2011). Still, it removes the requirement for external reference panels by building ancestral haplotypes from the data, similar to STITCH (Davies et al., 2016). The model reconstructs reference haplotypes dynamically and accounts for ancestry transitions caused by recombination events and miscopying errors using a Hidden Markov Model (HMM). LOCO identifies recombination switch points and estimates genome-wide ancestry proportions. To evaluate the performance of LOCO, we implemented a simulation framework that generates admixed genomes data under controlled parameters, including recombination rates, miscopying probabilities, and haplotype frequencies. We evaluate LOCO's ability in reconstructing diploid ancestry at 1000, 2000, and 4000 loci for sample sizes of 10, 20, and 40 individuals using the simulated admixed genomes data. LOCO correctly infers that 96.85–98.63% of diploid ancestry calls are made across different data scenarios. We also evaluated LOCO's ability to detect introgression events of different lengths. LOCO had difficulty recovering smaller introgressed segments, although it reliably found introgressed segments larger than roughly 75 loci with 93.38% of correct diploid ancestry calls. This study highlights LOCO's ability to infer local ancestry and detect introgression events when reference panels are unavailable, as well as by using genotype likelihoods obtained from low-depth sequencing data.

INTRODUCTION

Admixture exchanges genetic material between populations and creates human genetic diversity that contributes to differences in disease susceptibility across populations (Korunes & Goldberg, 2021). Studies using admixture mapping have shown loci associated with traits and diseases. These findings provide information on the genetic basis of complex traits (Dawkins & Lloyd, 2019; Divers et al., 2017; Garzón Rodríguez et al., 2024; Gomez et al., 2015; Koller et al., 2022; Pankratov et al., 2024; Welter et al., 2014; Xia et al., 2024). These studies show the importance of accurately detecting ancestry to understand the genetic architecture underlying phenotypic variation. Admixed individuals have a genome that is a mosaic of segments originating from two or more divergent populations. In diploid organisms, where each individual receives two sets of chromosomes, during meiosis, homologous chromosomal pairs exchange DNA

fragments through a process called crossing over (Kleckner, 2006). The crossing-over happens in certain locations known as chiasmata, which are the points of physical contact between two homologous chromosomes. The physical breaking and rejoining of DNA strands at these sites causes the alleles to be rearranged (Kleckner, 2006). Consequently, "switch points" are created where the ancestral origin changes from one population to another (Wegmann et al., 2011). In order to characterize admixture, it is necessary to 1) determine whether an individual is admixed, 2) map the genomic segments inherited from each ancestral group by locating recombination switch points, and 3) use reference panels to assign these segments to their respective populations. (Salter-Townshend & Myers, 2019). Reference panels comprise genotyping data from individuals with well-characterized and distinct ancestral backgrounds. Reference panels consist of selected genotyping data from individuals with well-defined ancestry (Salter-Townshend & Myers, 2019). Comparing the genetic markers in an admixed genome to those in the reference panel will identify the switch sites and determine the ancestry proportions of the genomic regions in the admixed individuals (Mosca & Cho, 2023; Salter-Townshend & Myers, 2019).

Probabilistic approaches are needed for ancestry detection because ancestral populations usually share alleles, given their shared evolutionary history. This overlap complicates the ancestry assignments to assign genetic contributions to one specific ancestral population (Shriner, 2013). Methods for ancestry inference are broadly categorized into allele-frequency-based and haplotype-based models (Padhukasahasram, 2014). Allele-frequency-based methods focus on how common alleles are in the reference populations. This approach was implemented in the LAMP (Local Ancestry in adMixed Populations) model, which determines the most likely ancestry at each locus using a majority voting system (Sankararaman et al., 2008).

LAMP requires no reference panel, which is an advantage in cases where ancestral populations of interest cannot be adequately sampled or when representative genetic data is unavailable. LAMP is highly accurate in differentiating between well-separated populations (for example, Africans and Europeans) compared to methods that rely on haplotype information. However, it does not perform well when the ancestral populations are very similar (Yuan et al., 2017). There is an extended version called LAMP-ANC,

which can improve accuracy by adding ancestral population data when available (Sankararaman et al., 2008). LAMP-ANC does not require a predefined reference panel, it uses ancestral population data derived directly from the study or other available sources. Other known allele-frequency-based methods are STRUCTURE and ADMIXTURE. The software STRUCTURE assigns individuals to populations and infers population structure using a Bayesian approach (Pritchard et al., 2000). It works even in cases where the actual population boundaries are unknown. The algorithm in the STRUCTURE framework calculates the likelihood that each individual's genome comes from one or more ancestral populations; the model uses Markov Chain Monte Carlo (MCMC) to estimate the allele frequencies for each population and explore various possible assignments(Pritchard et al., 2000). Similar to STRUCTURE, ADMIXTURE assigns ancestry proportions to individuals by modelling the data to estimate the allele **fr**equencies in each population without the need for established population labels, meaning it does not require prior knowledge of which individuals belong to which populations **(**Alexander et al., 2009). ADMIXTURE uses a maximum likelihood framework rather than a Bayesian. The Bayesian framework is scalable and fast, especially when working with big genome-wide datasets with thousands of samples. (Alexander et al., 2009).

In contrast to approaches that rely on allele frequencies, haplotype-based approaches commonly employ Hidden Markov Models (HMMs) to utilise haplotype structure for ancestry at a finer scale (Price et al., 2009; Shriner, 2013). This finer resolution comes when haplotype-based methods consider the contiguous segments of DNA inherited together from ancestors, allowing them to detect smaller, more localised ancestry contributions than allele-frequency-based methods (Guan, 2014). This approach was implemented in HAPMIX (HAPlotype MIXture), which models the ancestry process of assignment using phased reference panels (Price et al., 2009). This made it possible to identify small genomic fragments with different ancestral origins. The foundation of HAPMIX is the idea that the admixed population is descended from a combination of two ancestral populations. (Price et al., 2009). To determine the probability that a specific haplotype segment comes from one or more populations, HAPMIX uses phased

reference data. This methodology allows HAPMIX to detect admixture patterns by combining population-specific mutation rates and recombination maps and accounting for genotyping errors; all this information improves the robustness of the ancestry inferences. Recombination maps are a detailed representation of the frequency of recombination across the genome (Myers et al., 2005; Price et al., 2009). These maps identify regions where recombination occurs more frequently and those where it occurs less frequently. High-resolution recombination maps are useful in ancestry inference methods like HAPMIX because they enable the algorithm to detect the precise sites at which genetic ancestry switches between populations. This information improves the model's robustness by allowing it to better account for fine-scale structure in genetic data, especially in regions with high recombination rates and frequent transitions (Price et al., 2009). This information on recombination events is essential for accurately tracing ancestry contributions. The errors in modelling the transitions can lead to incorrect ancestry assignments and reduce the robustness of inference in admixed populations. (Hassan et al., 2021). The reliance of HAPMIX on reference panels and recombination maps restricts its application, particularly for populations with insufficiently characterised or missing reference data. These dependencies highlight the need for new methodologies to infer ancestry without such pre-requirements.

Methods like RASPberry were developed to get around the need for pre-established recombination maps. RASPberry analyses patterns of genetic variation and ancestry transitions; with this, it can derive recombination rates directly from genotype data (Wegmann et al., 2011). RASPberry uses an HMM framework to determine ancestry switch points by modelling several key parameters. The parameters estimated by RASPberry are: individual-specific ancestry proportions that indicate the probability of each segment originating from a given ancestral population; recombination of within and between-population events, reflecting the rates at which these transitions occur; copying probabilities to decide which reference haplotype is selected during a recombination event, and also incorporates miscopying probabilities to account for the possibility of switching to a haplotype from another population due to historical sharing or incomplete lineage sorting. RASPberry's strategy still requires accurate reference panels to determine the ancestry of chromosomal segments.

MOSAIC addresses another key limitation, *i.e.* the requirement for predefined relationships between reference panels and ancestral populations. In the HAPMIX model, the reference panels are assumed to closely match the genetic profiles of the ancestral populations they come from, which helps to have an accurate ancestry assignment. However, this assumption can be problematic in cases where the relationships between donor haplotypes, that are the specific genetic sequences representative of the ancestral populations they are derived from, and the true ancestral populations are uncertain or incomplete. To overcome this problem, MOSAIC employs nested Hidden Markov Models (HMMs) to infer these relationships directly from the data, rather than relying on predefined matches (Salter-Townshend & Myers, 2019). The HMM allows MOSAIC to determine dynamically how reference haplotypes relate to unseen ancestral groups and infer ancestry segments and admixture events without requiring prior knowledge of these relationships. As a result, MOSAIC can be applied effectively even when reference panels are imperfectly aligned (Salter-Townshend & Myers, 2019). While MOSAIC eliminates the reliance on predefined relationships between donor haplotypes and ancestral populations, the availability and quality of reference panels remain dependencies for robust ancestry assignments.

While admixture analysis relies on high-quality sequencing data -with sufficient coverage and low error rates- and reference panels, genotyping imputation is a separate but complementary field. In some genomic studies, the quality of raw sequencing data might be degraded by variables such as inadequate sequencing coverage, technical errors, or missing data due to genotyping platform restrictions (Browning & Browning, 2016). These issues result in incomplete genotyping datasets. There are genotype imputation tools focused on predicting missing genotypes. They achieve this through the use of a population's linkage disequilibrium structure, which is defined as the non-random association of alleles at distinct loci (Browning & Browning, 2016; Howie et al., 2012). Imputation methods use probabilistic frameworks to infer missing genetic data from observed patterns. For example, IMPUTE2 is a software based on an HMM framework that treats the unobserved true haplotypes as hidden states, the observed genotype data are generated from these hidden states, which represent the underlying haplotype structure from the well-characterized reference

panel (Howie et al., 2012). This model enables IMPUTE2 to accurately predict missing genotypes by "copying" segments from reference haplotypes, preserving the linkage disequilibrium patterns in the data. This approach has the advantage of high imputation accuracy when a reference panel is available, but its reliance on external reference panels can be a constraint for populations that do not have well-curated reference panels. Another software is BEAGLE, which uses localised haplotype clustering rather than a global HMM framework (Browning & Browning, 2016). BEAGLE detects clusters of similar haplotypes in small genomic regions by capturing the local linkage disequilibrium structure and then uses these clusters to predict missing genotypes based on the alleles found within them. This localised technique makes BEAGLE more computationally efficient than IMPUTE2 and independent of reference panels, relying only on accessible data. The disadvantage is that BEAGLE may provide less accuracy than IMPUTE2 where fine-scale haplotype structure is significant. Both IMPUTE2 and BEAGLE function best when genotype data have high call rates, low error rates, and sufficient marker density to capture linkage disequilibrium patterns across the genome. For situations where high-quality data and reference panels are limited STITCH is an alternative tool. STITCH (Sequencing To Imputation Through Constructing Haplotypes) addresses some of these challenges by offering a probabilistic framework that uses genotype likelihoods, derived from low-coverage sequencing data, to impute genotypes and infer haplotypes without requiring external reference panels (Davies et al., 2016). STITCH models each chromosome as a mosaic of K unknown ancestral haplotypes, after it infers the haplotype switch points directly from sequencing reads. These switch points are the transitions between inferred haplotypes; they are modeled probabilistically but are not explicitly associated with specific ancestral populations or ancestry transitions. STITCH effectively imputes genotypes in settings where traditional haplotype reference panels are unavailable, such as in non-human or genetically diverse populations. STITCH reduces dependency on reference panels, however, its primary focus is genotype imputation, not ancestry inference. STITCH lacks features critical for ancestry inference, such as explicitly modelling ancestry transitions, recombination events, or miscopying errors, which are essential for understanding the mosaic structure of admixed genomes. Also, STITCH does not assign ancestry to the

inferred haplotypes, limiting its use for studies requiring precise ancestry mapping. These limitations highlight the need for an alternative approach that can extend the capabilities of STITCH by not requiring predefined reference panels and using the sequencing data to provide the necessary inputs for ancestry inference.

In this study, we present a new ancestry inference model for admixed individuals that uses genotype likelihoods derived from sequencing data that are assumed to have low quality. Our low-depth copy model does not require reference panels; it uses its data to create reference haplotypes. We model ancestry transitions through recombination events and miscopying, treating each chromosome as a mosaic of segments that copy from the reconstructed reference haplotypes. We used C++ to build our approach and conducted simulations to verify it. By running the software, we created simulated data to predict ancestry after we inferred the known ancestry. Additionally, we assessed the model's detection limits of introgression by simulating introgression events and using our LOCO model to detect them.

1 Model

Consider I potentially admixed individuals with genetic data at L bi-allelic loci.

Following others Li et al. 2003, Price et al. 2009, Wegmann et al. 2011, Davies et al. 2016, we will model the chromosomes of these individuals as a mosaic of segments from N_p reference haploytpes for each of $p = 1, \ldots, P$ ancestral populations using a Li & Stepehens copy-model similar to `HapMix` and `RASPberry`, but infer reference haplotypes as proposed in `STITCH`.

1.1 Copy-Model

Let us denote by $\boldsymbol{z}_{il} = (z_{il}^{(1)}, z_{il}^{(2)})$ the pair of unknown haplotypes at individual i and locus l. These haplotypes copy from a particular reference haplotype $h_{pn}, n = 1, \ldots, N_p$, which may change along the chromosome due to recombination. Each haplotype $\boldsymbol{z}_i^{(k)} = (z_{i1}^{(k)}, \ldots, z_{iL}^{(k)}), k = 1, 2$ is thus given by a specific path through the reference haplotypes $h(z_{il}^{(k)}) \in \{h_{11}, \ldots, h_{1N_1}, \ldots, h_{PN_P}\}$, which we will model as follows:

1. Recombination events are modeled as a Poisson processes along the chromosome with rates scaled by the (genetic) distance δ_l between adjacent pairs of loci $l-1$ and l.
2. Recombination events within a population (i.e. that do not change the ancestry) occur with rates proportional to ρ_p.
3. Recombination events between populations (i.e. that may switch the ancestry) occur with rates proportional to ρ^*.
4. Recombination events resulting in a haplotype from population p pick haplotype $n = 1, \ldots, N_p$ with probability $f_{pn}, \sum_{n=1}^{N_p} f_{pn} = 1$. Note that thus a recombination event within population p results in a different haplotype with probability $1 - f_{pn}$.
5. A recombination event between populations occurring on a haplotype of individual i with current ancestry p' results in ancestry p with probability $\pi_{ip}, \sum_p \pi_{ip} = 1$, the genome-wide ancestry proportions of that individual. Note that thus a recombination event results in an ancestry switch with probability $1 - \pi_{ip}$.
6. Following Price et al. 2009, Wegmann et al. 2011, we will also allow for "miscopying" such that a recombination event resulting in population p picks a reference haplotype of that population with probability $\bar{q}_p = 1 - q_p$ and a haplotype of any other population with probability q_p. This process allows for historically shared haplotypes between populations, for instance, due to incomplete lineage sorting Price et al. 2009.

Let us denote the hidden state of the resulting HMM by the triplet $(p\tilde{p}n)$, indicating the ancestry $p = 1, \ldots, P$ and the haplotype $n = 1, \ldots, N_{\tilde{p}}$ copied from population $\tilde{p} = 1, \ldots, P$, where $\tilde{p}$ may differ from p due to miscopying. Let us further denote by $R_{lp} = 1 - e^{-\delta_l \rho_p}$ and $\bar{R}_{lp} = 1 - R_{lp}$, respectively, the probabilities that at least one and no recombination event occurred within population p between loci $l-1$ and l. Analogous, let $R_l^* = 1 - e^{-\delta_l \rho^*}$ and $\bar{R}_l^* = 1 - R_l^*$ denote, respectively, the probabilities that at least one and no between population recombination event occurred between loci $l-1$ and l.

The transition probabilities $z_{il-1}^{(k)} = (p'\tilde{p}'n') \rightarrow z_{il}^{(k)} = (p\tilde{p}n)$ are given by

$$\mathbb{P}\left(z_{il}^{(k)}|z_{il-1}^{(k)}\right) = \begin{cases} R_l^*\pi_{ip}\bar{q}_p f_{\tilde{p}n} & \text{if } p \neq p' \text{ and } p = \tilde{p} \quad \text{(A)} \\ R_l^*\pi_{ip} q_p f_{\tilde{p}n} & \text{if } p \neq p' \text{ and } p \neq \tilde{p} \quad \text{(A)} \\ \left(\bar{R}_l^* R_{lp} + R_l^*\pi_{ip}\right)\bar{q}_p f_{\tilde{p}n} & \text{if } p = p' \text{ and } p = \tilde{p} \text{ and } (\tilde{p} \neq \tilde{p}' \text{ or } n \neq n') \quad \text{(B)} \\ (\bar{R}_l^* R_{lp} + R_l^*\pi_{ip}) q_p f_{\tilde{p}n} & \text{if } p = p' \text{ and } p \neq \tilde{p} \text{ and } (\tilde{p} \neq \tilde{p}' \text{ or } n \neq n') \quad \text{(B)} \\ \bar{R}_l^*\bar{R}_{lp} + (\bar{R}_l^* R_{lp} + R_l^*\pi_{ip})\bar{q}_p f_{\tilde{p}n} & \text{if } p = p' \text{ and } p = \tilde{p} \text{ and } \tilde{p} = \tilde{p}' \text{ and } n = n' \quad \text{(C)} \\ \bar{R}_l^*\bar{R}_{lp} + (\bar{R}_l^* R_{lp} + R_l^*\pi_{ip}) q_p f_{\tilde{p}n} & \text{if } p = p' \text{ and } p \neq \tilde{p} \text{ and } \tilde{p} = \tilde{p}' \text{ and } n = n' \quad \text{(C)} \end{cases}$$

The probability for the initial state (first locus of a chromosome) $z_{i1}^{(k)}$ is given by

$$\mathbb{P}\left(z_{i1}^{(k)} = (p\tilde{p}n)\right) = \begin{cases} \pi_{ip}\bar{q}_p f_{\tilde{p}n} & \text{if } p = \tilde{p}, \quad \text{(a)} \\ \pi_{ip} q_p f_{\tilde{p}n} & \text{if } p \neq \tilde{p}. \quad \text{(b)} \end{cases}$$

1.2 Emission probabilities

Following (**Davies2016**), we will model each reference haplotype $\boldsymbol{h}_{pn}$ from population p as a vector of probabilities $\boldsymbol{h}_{pn} = (h_{pn1}, \ldots, h_{pnL})$ with which the haplotype emits an alternative allele at locus $l = 1, \ldots, L$.

Given the pair of haplotypes $\boldsymbol{z}_{il} = (z_{il}^{(1)}, z_{il}^{(2)})$ of individual i at locus l and denoting by $h_l(z_{il}^{(k)})$ the probability of the alternative allele at locus l of haplotype $h(z_{il}^{(k)})$, the probability of observing genotype $g_l = 0, 1, 2$ is

$$\mathbb{P}(g_l = g|\boldsymbol{z}_{il}) = \begin{cases} (1 - h_l(z_{il}^{(1)})(1 - h_l(z_{il}^{(2)})) & \text{if g} = 0, \quad \text{(I)} \\ (1 - h_l(z_{il}^{(1)}))h_l(z_{il}^{(2)}) + h_l(z_{il}^{(1)})(1 - h_l(z_{il}^{(2)})) & \text{if g} = 1, \quad \text{(II)} \\ h_l(z_{il}^{(1)})h_l(z_{il}^{(2)}) & \text{if g} = 2. \quad \text{(III)} \end{cases}$$

Given genetic data summarized by the genotype likelihoods $\mathbb{P}(d_{il}|g), g = 0, 1, 2$, we have

$$\mathbb{P}(d_{il}|z_{il}) = \sum_g \mathbb{P}(d_{il}|g)\mathbb{P}(g|z_{il}). \quad (1)$$

1.3 Hidden Markov Model

Let us define by $\boldsymbol{\theta} = \{\boldsymbol{\rho}, \rho^*, \boldsymbol{\pi}, \boldsymbol{q}, \boldsymbol{f}, \boldsymbol{h}\}$ the set of hierarchical parameters, where $\boldsymbol{\rho} = \rho_1, \ldots, \rho_P$ and $\boldsymbol{\pi} = \pi_{1,1}, \ldots \pi_{IP}$ and $\boldsymbol{q} = q_1, \ldots, q_P$ and $\boldsymbol{f} = f_{1,1}, \ldots, f_{PN_P}$ and $\boldsymbol{h} = h_{111}, \ldots, h_{PN_PL}$. In addition, let us define by $\boldsymbol{z} = z_{11}^{(k)}, \ldots, z_{IL}^{(k)}$ all hidden states and by $\mathcal{D} = d_{11}, \ldots, d_{IL}$ all observed data. The complete data likelihood is given then by

$$\mathcal{L}_c(\boldsymbol{\theta}) = \mathbb{P}(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{z}) = \prod_{i=1}^{I} \mathbb{P}(\boldsymbol{z}_{i1}) \prod_{l=2}^{L} \mathbb{P}(\boldsymbol{z}_{il}|\boldsymbol{z}_{il-1}) \prod_{l=1}^{L} \mathbb{P}(d_{il}|\boldsymbol{z}_{il}), \quad (2)$$

where the initial state and transition probabilities are given by the product over the two haplotypes:

$$\begin{aligned} \mathbb{P}(\boldsymbol{z}_{i1}) &= \mathbb{P}(z_{i1}^{(1)})\mathbb{P}(z_{i1}^{(2)}), \\ \mathbb{P}(\boldsymbol{z}_{il}|\boldsymbol{z}_{il-1}) &= \mathbb{P}(z_{il}^{(1)}|z_{il-1}^{(1)})\mathbb{P}(z_{il}^{(2)}|z_{il-1}^{(2)}). \end{aligned}$$

Note that we chose a notation with one HMM running on the combination of the two hidden states $\boldsymbol{z}_{il}$, rather than two HMMs running on each of them independently, because this simplifies the notation. Mathematically, the two options are equivalent.

The complete data log-likelihood is

$$\ell_c(\boldsymbol{\theta}) = \log \mathbb{P}(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{z}) = \sum_{i=1}^{I} \log \mathbb{P}(\boldsymbol{z}_{i1}) + \sum_{i=1}^{I}\sum_{l=2}^{L} \log \mathbb{P}(\boldsymbol{z}_{il}|\boldsymbol{z}_{il-1}) + \sum_{i=1}^{I}\sum_{l=1}^{L} \log \mathbb{P}(d_{il}|\boldsymbol{z}_{il}). \tag{3}$$

1.3.1 Q-function

Maximum-likelihood estimates of $\boldsymbol{\theta}$ can be obtained by iteratively maximizing the Q-function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathbb{E}\left[\ell_c(\boldsymbol{\theta})|\boldsymbol{\theta}', \boldsymbol{d}_{1:L}\right]$, where $\boldsymbol{\theta}'$ denotes the vector of current parameter estimates:

$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}') &= \sum_{i=1}^{I}\Bigg[\mathbb{E}\left[\log \mathbb{P}(\boldsymbol{z}_{i1})|\boldsymbol{d}_{i,1:L}, \boldsymbol{\theta}'\right] + \sum_{l=2}^{L}\mathbb{E}\left[\log \mathbb{P}(\boldsymbol{z}_{il}|\boldsymbol{z}_{il-1})|\boldsymbol{d}_{i,1:L}, \boldsymbol{\theta}'\right] + \dots \\
&\quad + \sum_{l=1}^{L}\mathbb{E}\left[\log \mathbb{P}(d_{il}|\boldsymbol{z}_{il})|\boldsymbol{\theta}'\right]\Bigg] \\
&= \sum_{i=1}^{I}\Bigg[\sum_{\boldsymbol{z}_{i1}}\gamma_{i1}(\boldsymbol{z}_{i1})\log \mathbb{P}(\boldsymbol{z}_{i1}) + \sum_{l=2}^{L}\sum_{\boldsymbol{z}_{il}}\sum_{\boldsymbol{z}_{il-1}}\xi_{il}(\boldsymbol{z}_{il}, \boldsymbol{z}_{il-1})\log \mathbb{P}(\boldsymbol{z}_{il}|\boldsymbol{z}_{il-1}) + \dots \\
&\quad + \sum_{l=1}^{L}\sum_{\boldsymbol{z}_{il}}\gamma_{il}(\boldsymbol{z}_{il})\log \mathbb{P}(d_{il}|\boldsymbol{z}_{il})\Bigg],
\end{aligned}$$

where $\gamma_{il}(\boldsymbol{z}_{il})$ and $\xi_{il}(\boldsymbol{z}_{il}, \boldsymbol{z}_{il-1})$ denote the expectation weights $\mathbb{P}(\boldsymbol{z}_{il}|\boldsymbol{d}_{i,1:L})$ and $\mathbb{P}(\boldsymbol{z}_{il}, \boldsymbol{z}_{il-1}|\boldsymbol{d}_{i,1:L})$, respectively. These weights can be calculated efficiently with the standard forward-backward algorithm.

2 Inference

2.1 Parameters of the transition-part

For the updates of the parameters in the EM, we need to take the derivative of the Q-function with respect to each parameter and solve for zero. Since this is analytically not feasible, we will resort a Newton-Raphson to find the root of first derivative, and hence also require the second derivatives.

To unclutter the notation, let us re-arrange the sums in the transition-part of the Q-function:

$$\begin{aligned}
Q_t(\boldsymbol{\theta}|\boldsymbol{\theta}') &= \sum_{i=1}^{I}\sum_{l=2}^{L}\sum_{\boldsymbol{z}_{il}}\sum_{\boldsymbol{z}_{il-1}}\xi_{il}(\boldsymbol{z}_{il}, \boldsymbol{z}_{il-1})\log \mathbb{P}(\boldsymbol{z}_{il}|\boldsymbol{z}_{il-1}), \\
&= \sum_{i=1}^{I}\sum_{l=2}^{L}\sum_{z_{il}^{(1)}}\sum_{z_{il}^{(2)}}\sum_{z_{il-1}^{(1)}}\sum_{z_{il-1}^{(2)}}\xi_{il}(z_{il}^{(1)}, z_{il}^{(2)}, z_{il-1}^{(1)}, z_{il-1}^{(2)})\left(\log \mathbb{P}(z_{il}^{(1)}|z_{il-1}^{(1)}) + \log \mathbb{P}(z_{il}^{(2)}|z_{il-1}^{(2)})\right), \\
&= \sum_{i=1}^{I}\sum_{l=2}^{L}\Bigg[\sum_{z_{il}^{(1)}}\sum_{z_{il-1}^{(1)}}\log \mathbb{P}(z_{il}^{(1)}|z_{il-1}^{(1)})\sum_{z_{il}^{(2)}}\sum_{z_{il-1}^{(2)}}\xi_{il}(z_{il}^{(1)}, z_{il}^{(2)}, z_{il-1}^{(1)}, z_{il-1}^{(2)}) + \dots \\
&\quad + \sum_{z_{il}^{(2)}}\sum_{z_{il-1}^{(2)}}\log \mathbb{P}(z_{il}^{(2)}|z_{il-1}^{(2)})\sum_{z_{il}^{(1)}}\sum_{z_{il-1}^{(1)}}\xi_{il}(z_{il}^{(1)}, z_{il}^{(2)}, z_{il-1}^{(1)}, z_{il-1}^{(2)})\Bigg].
\end{aligned}$$

By summarizing by transition case $X \in \{A, B, C\}$ and current state $(p\tilde{p}n)$, we arrive at

$$Q_t(\boldsymbol{\theta}) = \sum_i \sum_l \sum_X \sum_{p\tilde{p}n} \xi_{il}^{(X)_{p\tilde{p}n}} \log\left(q_{il}^{(X)_{p\tilde{p}n}}(\boldsymbol{\theta})\right),$$

where $\xi_{il}^{(X)_{p\tilde{p}n}}$ sums over all entries of the ξ_{il} matrix where haplotype (1) or (2) take transition X and end up in current state $(p\tilde{p}n)$:

$$\begin{aligned}\xi_{il}^{(X)_{p\tilde{p}n}} &= \sum_{z_{il}^{(1)}} \mathcal{I}\left(z_{il}^{(1)} = (p\tilde{p}n)\right) \sum_{z_{il-1}^{(1)}} \mathcal{I}\left((z_{il-1}^{(1)} \to z_{il}^{(1)}) \in X\right) \sum_{z_{il}^{(2)}} \sum_{z_{il-1}^{(2)}} \xi_{il}(z_{il}^{(1)}, z_{il}^{(2)}, z_{il-1}^{(1)}, z_{il-1}^{(2)}) + \\ &\sum_{z_{il}^{(2)}} \mathcal{I}\left(z_{il}^{(2)} = (p\tilde{p}n)\right) \sum_{z_{il-1}^{(2)}} \mathcal{I}\left((z_{il-1}^{(2)} \to z_{il}^{(2)}) \in X\right) \sum_{z_{il}^{(1)}} \sum_{z_{il-1}^{(1)}} \xi_{il}(z_{il}^{(1)}, z_{il}^{(2)}, z_{il-1}^{(1)}, z_{il-1}^{(2)}).\end{aligned}$$

Here, $\mathcal{I}\left((z_{il-1}^{(k)} \to z_{il}^{(k)}) \in X\right)$ indicates if the transition from $z_{il-1}^{(k)}$ to $z_{il}^{(k)}$ belongs to the group of transitions characterized in transition X, and $\mathcal{I}\left(z_{il}^{(k)} = (p\tilde{p}n)\right)$ indicates if the state $z_{il}^{(k)}$ corresponds to the triplet $(p\tilde{p}n)$. For each transition case $X \in A, B, C$, we therefore store a total of $\sum_p \sum_{\tilde{p}} N_{\tilde{p}}$ sums.

2.2 Parameters of the initial-part

For simplicity, we omit the initial probabilities when updating parameters.

2.3 Parameters of the emission-part

For the emission probabilities, the emission-part of the Q-function is relevant, i.e. $Q_e(\boldsymbol{\theta}|\boldsymbol{\theta}')$ that was written out above.

$$Q_e(\boldsymbol{\theta}|\boldsymbol{\theta}') = \sum_{i=1}^{I} \left[\sum_{l=1}^{L} \sum_{\boldsymbol{z}_{il}} \gamma_{il}(\boldsymbol{z}_{il}) \log \mathbb{P}(d_{il}|\boldsymbol{z}_{il})\right], \tag{4}$$

Given the genotype likelihoods $\mathbb{P}(d_{il}|g)$

$$Q_e(\boldsymbol{\theta}|\boldsymbol{\theta}') = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{\boldsymbol{z}_{il}} \gamma_{il}(\boldsymbol{z}_{il}) \log\left(\sum_g \mathbb{P}(d_{il}|g)\mathbb{P}(g|\boldsymbol{z}_{il}))\right), \tag{5}$$

where

$$\begin{aligned}\sum_g \mathbb{P}(d_{il}|g)\mathbb{P}(g|\boldsymbol{z}_{il})) &= \mathbb{P}(d_{il}|g=0)(1-h_l(z_{il}^{(1)})(1-h_l(z_{il}^{(2)})) + \ldots \\ &+\mathbb{P}(d_{il}|g=1)((1-h_l(z_{il}^{(1)}))h_l(z_{il}^{(2)}) + h_l(z_{il}^{(1)})(1-h_l(z_{il}^{(2)}))) + \ldots \\ &+\mathbb{P}(d_{il}|g=2)h_l(z_{il}^{(1)})h_l(z_{il}^{(2)}).\end{aligned}$$

take derivative with respect to $h_l(z_{il}^{(k)}), k = 1, 2$. Let us define by $h_l(z_{il}^{(j)})$ the h_l for the $j \neq k$ (i.e. when deriving for $h_l(z_{il}^{(1)})$, then $j = 2$, and vice versa)

$$\frac{\partial}{\partial h_l(z_{il}^{(k)})} Q_e(\boldsymbol{\theta}|\boldsymbol{\theta}') = \sum_{i=1}^{I} \sum_{z_{il}^{(j)}} \gamma_{il}(z_{il}^{(k)}, z_{il}^{(j)}) \left(\frac{S(j)}{G}\right), \tag{6}$$

where $G = \sum_g \mathbb{P}(d_{il}|g)\mathbb{P}(g|z_{il}^{(k)}, z_{il}^{(j)})$ and

$$S(s) = -\mathbb{P}(d_{il}|g=0)(1-h_l(z_{il}^{(s)})) + \mathbb{P}(d_{il}|g=1)(1-2h_l(z_{il}^{(s)})) + \mathbb{P}(d_{il}|g=2)h_l(z_{il}^{(s)}).$$

Set to zero and solve for $h_l(z_{il}^{(k)})$ is analytically not feasible. We will employ Newton-Raphson and the second derivatives are required.

$$\frac{\partial}{\partial h_l(z_{il}^{(k)})^2} Q_e(\boldsymbol{\theta}|\boldsymbol{\theta}') = \sum_{i=1}^{I} \sum_{z_{il}^{(j)}} \gamma_{il}(z_{il}^{(k)}, z_{il}^{(j)}) \left(\frac{-S(j)S(k)}{G^2} \right). \tag{7}$$

and

$$\frac{\partial}{\partial h_l(z_{il}^{(k)}) h_l(z_{il}^{(j)})} Q_e(\boldsymbol{\theta}|\boldsymbol{\theta}') = \sum_{i=1}^{I} \gamma_{il}(z_{il}^{(k)}, z_{il}^{(j)}) \left(\frac{\left(\mathbb{P}(d_{il}|g=0) - 2\mathbb{P}(d_{il}|g=1) + \mathbb{P}(d_{il}|g=2)\right) G - S(j)S(k)}{G^2} \right), \tag{8}$$

Implementation

We implemented the proposed HMM framework as a C++ program called LOCO, which stands for LowDephCopyModel. The source code and documentation are publicly available through a git repository at https://bitbucket.org/wegmannlab/lowdepthcopymodel. LOCO is designed to be user-friendly, providing a command-line interface that allows users to specify input files. The program supports BEAGLE files to represent genotype likelihood data.

Simulation framework

To evaluate the model's performance, we implemented a simulation framework that generates admixed genomes using a probabilistic approach. Admixed individuals are generated using parameters representing global (rho_start) and population-specific recombination rates (rho_p), miscopying probabilities (q_p), population proportions (pi_ip), and haplotype frequencies (f_pn). The parameters rho_start, rho_p and q_p are provided via command-line input, and pi_ip and f_pn are sampled from a Dirichlet distribution where pi_ip values sum to 1 for each individual across population and for f_pn sum to 1 for each population across haplotypes. Simulations start by generating the hidden ancestry states for each individual across all loci using a transition probability matrix. This matrix represents the recombination events and miscopying and is computed based on the sample parameters rho_start, rho_p, q_p, pi_ip and f_pn. For each individual, the HMM simulates a sequence of hidden states corresponding to the haplotypes inherited from the ancestral populations. Once the hidden states are generated, they simulate genotype data. At each locus, the two haplotypes define the probability of the individual carrying 0, 1, or 2 alternative alleles. The probabilities are computed as follows, given the alternative allele probabilities θ_1 and θ_2 from each haplotype:

- $P(G=0) = (1-\theta_1)(1-\theta_2)$
- $P(G=1) = (1-\theta_1)\theta_2 + \theta_1(1-\theta_2)$
- $P(G=2) = \theta_1\theta_2$

Genotypes are sampled according to these probabilities and written in a Beagle-format output file. Genotype likelihoods are hard-called, meaning the selected genotype has probability 1, and the others have probability 0. We initialise the model with values close to the true simulated parameters with an error of 0.0001 for inference.

Simulation of ancestry

To generate the simulated data used to evaluate the performance of the ancestry inference. We executed the simulation framework as follows: combination runs using the number of individuals 10, 20 and 40, and the number of loci 1000, 2000 and 4000. For each simulation, we specified the number of populations 2, the number of haplotypes per population 2,2, rho_star=0.01, rho_p=0.1,0.2 and q_p=0, pi_ip and f_pnl sampled with variances of 0.1, h_pnl samples with shape parameters $\alpha = \beta = 0.7$. The output of each simulation run was written in Beagle format. For inference, the generated Beagle file and simulation parameters file were used as input. The haplotypes per population were fixed at 2,2, and the model was initialised with values close to the true simulation parameters, perturbed by 0.0001. Convergence was controlled with a maximum of 3000 iterations and a minimum log-likelihood change threshold of 0.0001.

Simulation of introgression

To evaluate the model's ability to detect introgressed genomic segments, we extended the simulation framework to simulate introgression events explicitly. This was done by introducing a donor population (Population 2) contributing a genomic segment to a subset of individuals in the simulated dataset. We simulated datasets with 10 individuals, 2000 loci, number of populations 3 and haplotypes per population 2,2,2. Individual 0 was designated as the donor, with 100% ancestry from Population 1. All other individuals were initialised with 0% ancestry from Population 1; this ensures that any observed ancestry from Population 1 results from the introgression. Introgression was simulated by forcibly overwriting a contiguous genomic segment with haplotypes from the donor population. The introgression region covers loci [1000, END], where END ranges from 1001 to 2000, simulating different introgressed segment lengths. For each simulation, there is a Beagle-format file and an information file containing the

information of which individuals received introgression. Inference was then performed using the same simulation framework. The number of haplotypes per population was fixed at 2,2,2, and the inference was initialised using the true simulation parameters and perturbed by 0.0001. Convergence was controlled with a maximum of 3000 iterations and a minimum log-likelihood change threshold of 0.0001.

Extraction of the population of origin

To get the population of origin, we extracted ancestry calls from the posterior distribution over all combined hidden states. Each state represents a diploid configuration of two haplotypes encoded as (p1,p_tilde1,n1)×(p2,p_tilde2,n2). To determine the true ancestral source, we marginalised the posterior distribution by focusing on each state's p1 and p2 components, which are the actual population from which haplotype 1 and haplotype 2 were inherited. To extract the diploid ancestry calls, we listed all possible single-haplotype states based on the number of populations and haplotypes per population. Then, for each individual and each loci, we used the posterior probabilities to build a population-by-population matrix, where each entry (i,j) represents the total probability that haplotypes 1 and 2 came from populations i and j, respectively. For example, this gave us the diploid ancestry call for two populations: P1/P1, P1/P2, or P2/P2 for every individual at each locus.

Diploid ancestry matching genotype calls

To evaluate the model's performance in recovering individuals' diploid ancestry, we compared the inferred ancestry calls to the true simulated ancestry. The comparison was done for the combination of individuals: 10, 20, 40 and loci: 1000, 2000, 4000. We used the posterior state probabilities output by the simulator and the inference model to derive diploid ancestry calls as done in the previous section of extraction of the population of origin. To quantify the percentage of matching genotype calls between simulated and inferred diploid ancestry calls, we computed the percentage of loci where the simulated and inferred calls matched each individual. This percentage represents the diploid ancestry matching genotype calls for that individual. The matching genotype

calls distributions were visualised across the different data sets, which are the combination of loci and individuals, using ridgeline plots.

Data availability statement

The outcomes of the simulations will be made available. For the submission of this thesis, the simulation results are provided in a separate folder. The code used for this study is part of the LowDepthCopyModel project and is currently hosted in a private Git repository on Bitbucket.

Evaluation of the Ancestry Inference

We compared the true simulated ancestry with the inferred ancestry to evaluate our model in inferring diploid ancestry across admixed genomes. The ancestry of each individual at each locus was represented by one of three possible diploid ancestry states: P1/P1 (homozygous), P1/P2(heterozygous), or P2/P2 (homozygous), which correspond to the combinations of haplotypes inherited from two ancestral populations. The inferred ancestry profiles for the dataset of 10 individuals and 1000 genomic loci captured the overall patterns of population contributions and ancestry transitions (Figure 1A). The total number of diploid ancestry calls was 10000, one for each locus in each individual. Of these, 9823 ancestry calls (98.23%) matched correctly between the simulated and inferred calls. Minor discrepancies were observed in 177 ancestry calls (1.77%). We identified 66 ancestry discrepancy segments across the 10 individuals. These segments varied in length from 1 to 14 loci, with most discrepancies being short, with a mean length of 2.68 loci. These discrepancies occurred around recombination breakpoints, where the ancestry changes rapidly from one population to another, and the true ancestral segments are very short, making them harder to detect. We increased the dataset size to 20 individuals and 2000 genomic loci (Figure 1B). The model continued to accurately recover the simulated diploid ancestry across the majority of loci (Figure 1B). The total number of diploid ancestry calls was 40000, one for each locus in each individual. These 39350 calls (98.375%) were consistent between the simulated and inferred ancestry. 650 calls (1.625%) did not match the simulated ancestry; We identified 220 ancestry discrepancy segments across the 20

individuals. These segments ranged from 1 to 34 loci, with most being short, with a mean of 2.95 loci. As in the smaller dataset, the discrepancies were generally localised and often occurred near recombination breakpoints, where the ancestry switches rapidly from one population to another. In such regions, the true ancestral segments can be very short, which makes them more difficult to resolve.

Regardless of the increase in sample size and genomic complexity, the model could reconstruct the broad patterns of ancestry transitions. We also computed the percentage of correctly matching calls of simulated and inferred diploid ancestry calls for combining data sets of 10, 20 and 40 individuals and 1000,2000 and 4000 loci (Figure 2). Overall, the mean matching percentages remained high, ranging from 96.85% to 98.63%, with the highest matching observed in the dataset of 20 individuals and 1000 loci (98.63%) and the lowest in the dataset of 10 individuals and 2000 loci (96.86%). Increasing the number of loci from 1000 to 4000 and the number of individuals from 10 to 40 did not compromise the matching percentage, underscoring the robustness of the model. The discrepancies increased in larger data sets, ranging from 177 discrepancies in the smallest dataset (10 individuals and 1000 loci) to 2402 in the largest (40 individuals and 4000 loci). The mean segment lengths of the discrepancies kept short ranged from 2.49 (20 individuals and 1000 loci) to 5.38 loci (10 individuals and 2000 loci). The model can recover diploid ancestry in independent inference runs with data sets of different genomic lengths and sample sizes.

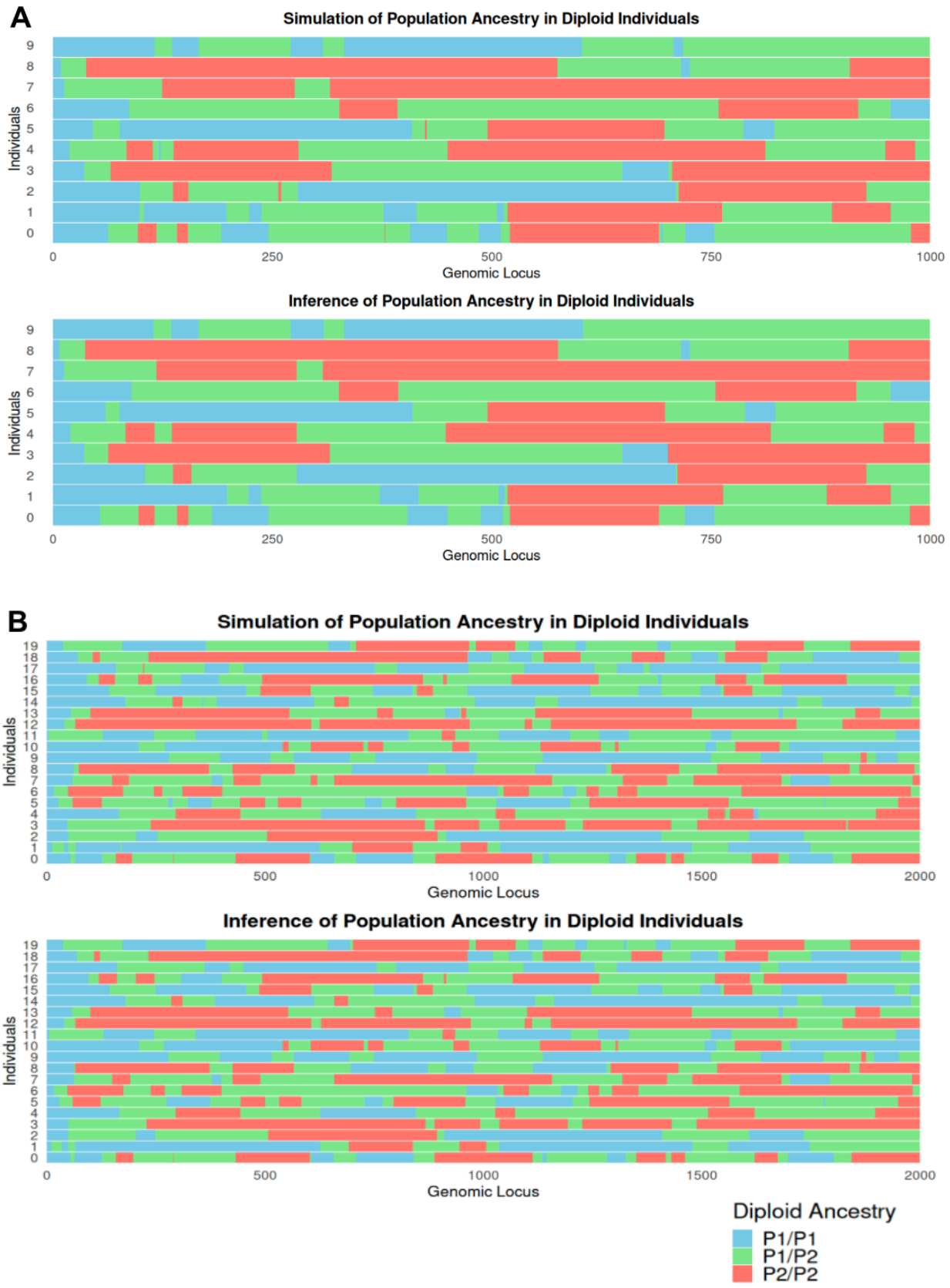
A
Simulation of Population Ancestry in Diploid Individuals
Individuals
Genomic Locus
Inference of Population Ancestry in Diploid Individuals
Individuals
Genomic Locus
B
Simulation of Population Ancestry in Diploid Individuals
Individuals
Genomic Locus
Inference of Population Ancestry in Diploid Individuals
Individuals
Genomic Locus
Diploid Ancestry
P1/P1
P1/P2
P2/P2

Figure 1. Comparison of Simulated and Inferred Diploid Ancestry. (A) Comparison for 10 individuals across 1000 genomic loci. (B) Comparison for 20 individuals across 2000 genomic loci. For each (A) and (B), the top panel shows the simulated ancestry for each individual at each locus, while the bottom panel shows the ancestry inferred by the LOCO model. Each colour represents a diploid ancestry state: P1/P1, P1/P2, and P2/P2.

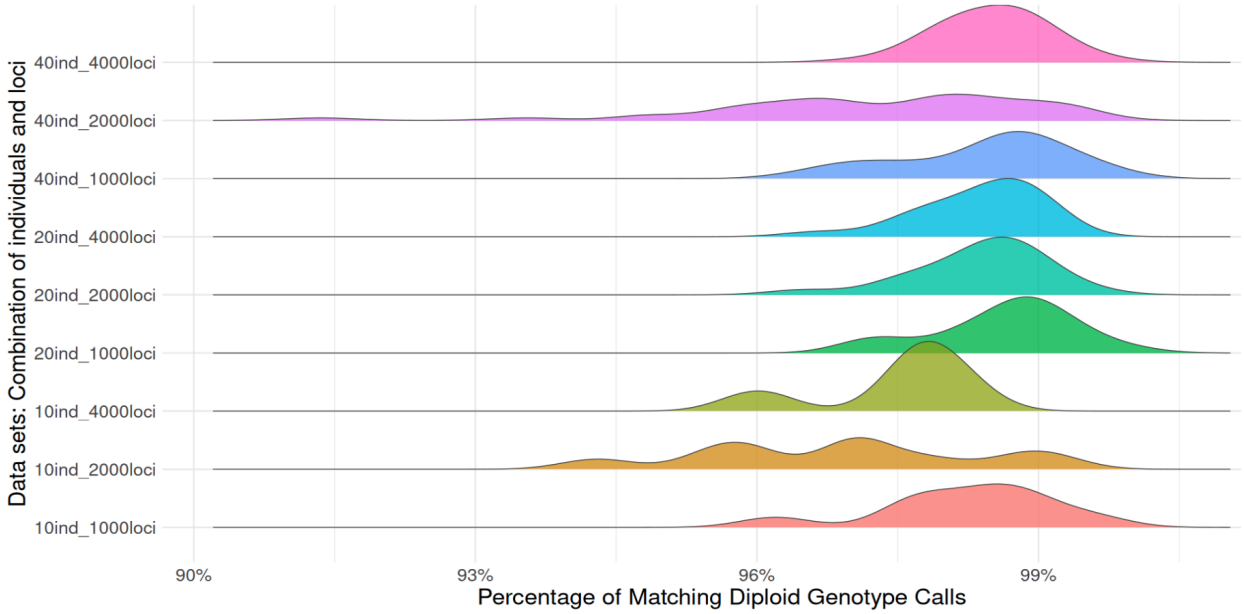

Figure 2. Percentage of Matched Diploid Genotype Calls Between Simulated and Inferred Ancestry. Ridgeline plots showing the distribution of per-individual diploid ancestry matching percentage across nine simulated datasets. Each dataset corresponds to a combination of the number of individuals (10, 20, or 40) and the number of genomic loci (1000, 2000, or 4000). For each individual, the matching percentage was calculated as the fraction of loci where the inferred diploid ancestry call (P1/P1, P1/P2, or P2/P2) matched the true simulated call.

Evaluation of the simulated introgression

To evaluate the model's ability to detect introgressed genomic segments, we simulated a dataset containing a short introgression event of 20 loci, originating from individual 0 as the donor, who carries ancestry P2/P2 (Figure 3A). The recipients of the introgression were individuals 1 to 4. In the simulation, we observe that the introgressed

segment was correctly introduced in individuals 1 to 4, showing ancestry P2/P2 in loci 1000–1020. However, in the inference, the 20 ancestry calls within the introgressed segment failed to match the simulated P2/P2 calls in these individuals. Instead, the inferred ancestry in this region consisted of P1/P3. This indicates that the model is unable to correctly infer the small introgressed segment of 20 loci. To evaluate the model's performance in detecting a longer introgressed genomic segment, we simulated a dataset containing an introgression event of 100 loci, originating from individual 0 as the donor, who carries ancestry P2/P2 (Figure 3B). The recipients of the introgression were individuals 1 to 4. In the simulation, the introgressed segment was correctly introduced in these individuals, showing ancestry P2/P2 in loci 1000–1100. In the inference, the model successfully recovered the introgressed segment in all four individuals. For individuals 2 to 4, all 100 loci within the introgressed region were correctly inferred as P2/P2. For individual 1, the model recovered 95 out of 100 loci accurately. The model is capable of correctly detecting longer introgressed segments.

To evaluate the model's sensitivity to introgressed segment length, we simulated introgression events of sizes from 1 to 1000 and evaluated the percentage of correctly inferred diploid ancestry calls. We focused on four recipient individuals, 1 to 4, and measured the percentage of matching ancestry calls between simulated and inferred segments as a function of introgression length (Figure 5). For short introgressed segments (0–25 loci), the model failed to recover any ancestry correctly, with an average matching percentage of 0%. In the 25–50 introgressed length range, the average matching percentage was 12.95%. In the 50–75 range, we have an average of 60.29%. For segments between 75 and 100 loci, the average matching percentage reached 93.38%. The smallest introgression segment length that resulted in a perfect 100% matching ancestry percentage was 41 loci introgressed in individual 3. Conversely, the largest introgressed segment that the model completely failed to detect was 80 loci in individual 4. The model has limited power to detect very short introgressed segments (<50 loci), but performs robustly for segments longer than 75 loci, with near complete recovery of true ancestry.

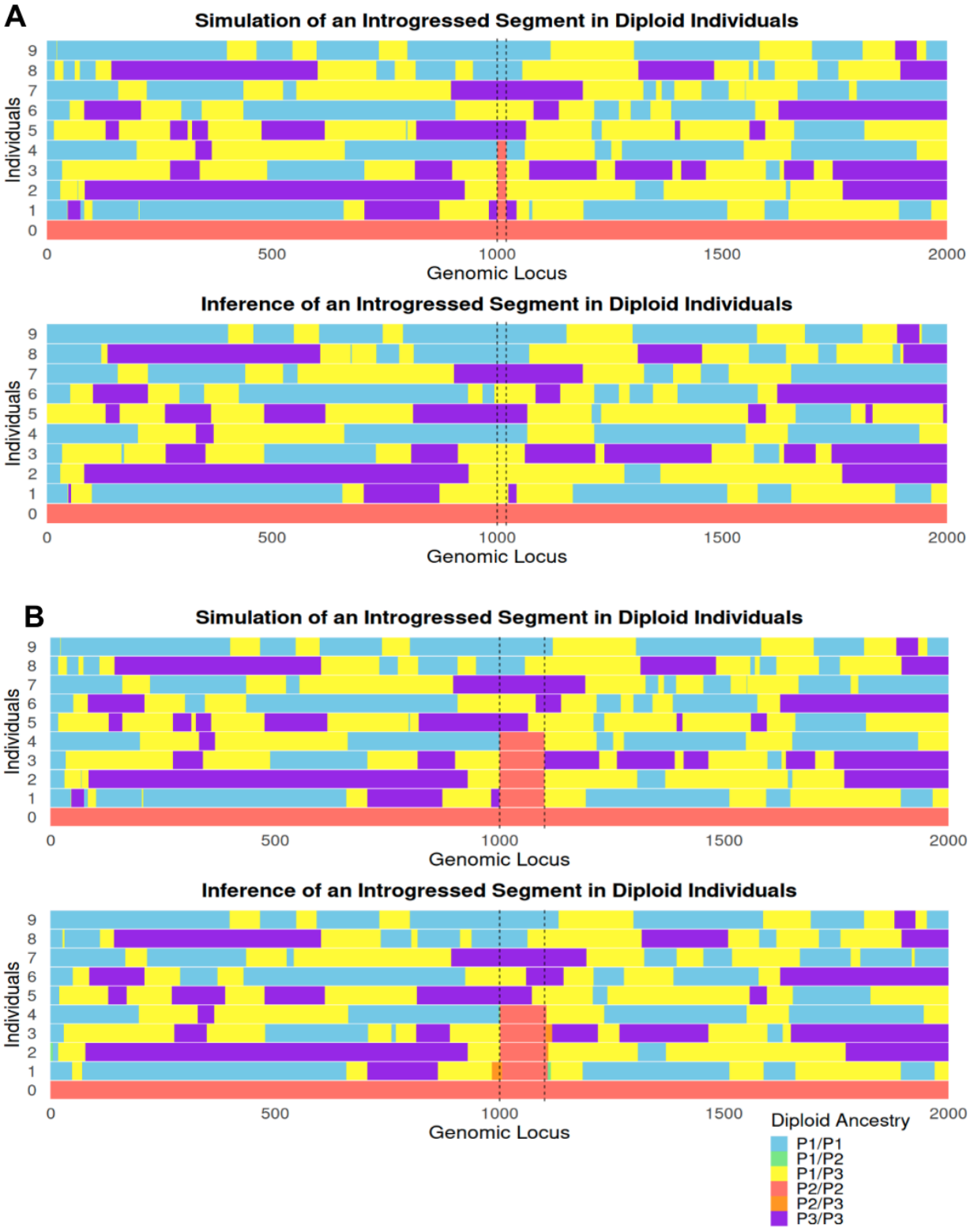
A
Simulation of an Introgressed Segment in Diploid Individuals
Individuals
Genomic Locus
Inference of an Introgressed Segment in Diploid Individuals
B
Simulation of an Introgressed Segment in Diploid Individuals
Inference of an Introgressed Segment in Diploid Individuals
Diploid Ancestry
P1/P1
P1/P2
P1/P3
P2/P2
P2/P3
P3/P3

Figure 3. Simulation and Inference of Introgressed Segment. (A) A short introgressed segment of 20 loci from positions 1000 to 1020. (B) A longer introgressed segment of 100 loci from positions 1000 to 1100. For each (A) and (B), the top panel shows the simulated ancestry with the introgressed segment originating from individual 0 as the donor (P2/P2), introduced into individuals 1 to 4. The bottom panel shows the ancestry inferred by the LOCO model. Each colour represents a diploid ancestry state: P1/P1, P1/P2, P1/P3, P2/P2, P2/P3, and P3/P3. The dashed lines represent the segment where the introgression can be found.

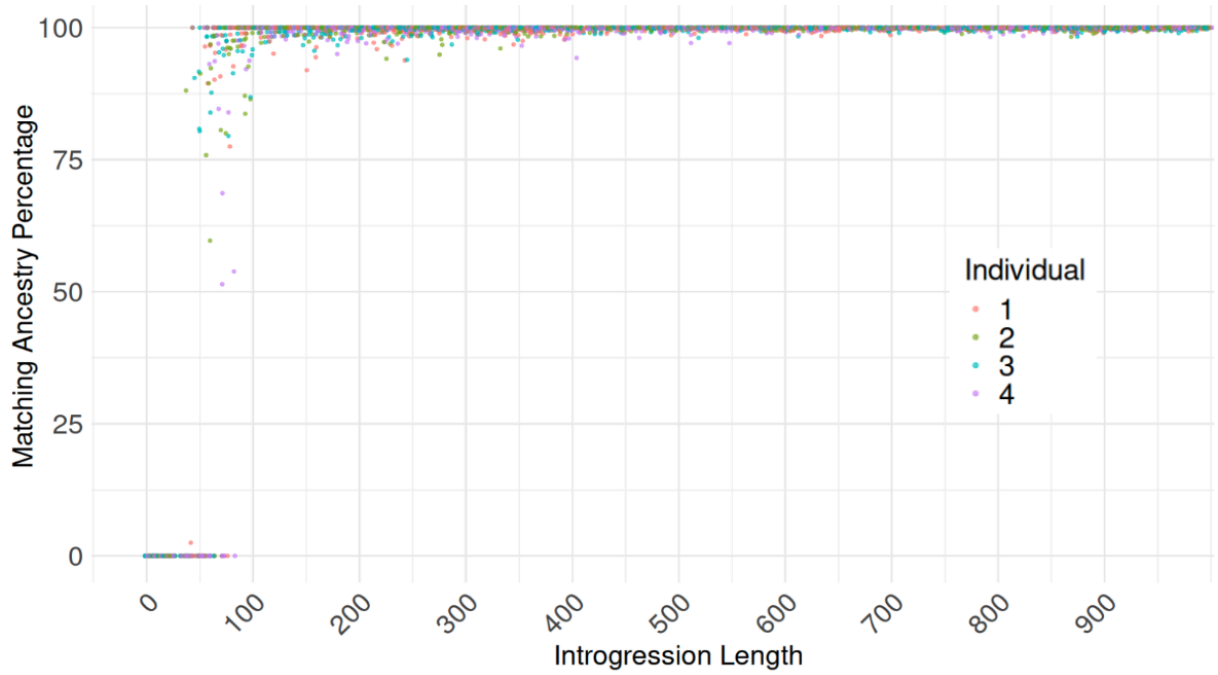

Figure 4. Introgression Length on Ancestry Inference. Each point represents the percentage of correctly inferred ancestry calls for an individual across an introgressed segment of a given length. The x-axis shows the length of the introgressed segment in loci, and the y-axis shows the percentage of matching ancestry calls between the simulated and inferred segments.

Discussion

In this study, we introduced a new Low-depth copy model to infer ancestry in admixed individuals directly from genotype likelihoods derived from low-quality sequencing data. Our approach builds upon Li & Stephens copy models, as extended by HAPMIX and RASPberry, and adopts the key feature from STITCH of constructing reference haplotypes from the data itself rather than requiring curated panels (N. Li & Stephens, 2003; Price et al., 2009; Wegmann et al., 2011). LOCO combines this data-driven haplotype reconstruction with explicit modeling of ancestry transitions, recombination events, and miscopying directly from genotype likelihoods within an HMM framework, allowing accurate ancestry inference in the absence of reference panels.

Our simulation results demonstrated correct performance in the inference of diploid ancestry calls across a range of genomic lengths from 1000, 2000 and 4000 loci and sample sizes of 10, 20 and 40 individuals. The percentage of matching diploid ancestry calls between the simulated and inferred data remained high in all scenarios, with mean values ranging from 96.85% to 98.63%. These results reflect the robustness of the model, even as genomic complexity and sample size increased. Discrepancies between the simulated and inferred ancestries happened at specific loci corresponding to recombination breakpoints, where ancestry changes abruptly from one population to another. Across all datasets, the discrepancies were segment lengths consistently short between 2.49 and 5.38 loci on average. These short segments were difficult for LOCO to recover, as their signal was indistinguishable from adjacent regions. As a result, LOCO tended to smooth over these transitions, assigning the dominant neighboring ancestry state across the region. This limitation of not correctly predicting the ancestry for short segments near breakpoints has been observed in other local ancestry models (Guan, 2014; Maples et al., 2013). Maples *et al.* showed that window-based discriminative methods, like RFMix, often exhibit switch errors at recombination boundaries because the signal in a few SNPs is insufficient to support a state change (Maples et al., 2013). Similarly, Guan *et al.* reported that where the ancestral track is short, it tends to exhibit lower inference accuracy due to insufficient distinguishing information (Guan, 2014). While LOCO effectively reconstructs broad ancestry profiles

and long-range transitions, it shares a common limitation with other local ancestry methods in recovering very short segments near recombination breakpoints. We further evaluated LOCO's ability to detect introgressed genomic segments. We observed that segments shorter than 25 loci had a 0% average matched to the correct introgressed ancestry, while those in the 25–50 loci range had only 12.95% matching. The performance improved significantly for longer segments, reaching an average of 93.38% matching for segments 75–100 loci in length. These findings indicate that LOCO is capable of detecting longer introgression segments, but its resolution for detecting short introgressed segments is limited. As mentioned above, having short segments in this case from introgression events makes it challenging for inference models to differentiate these segments from the surrounding genomic background (Guan, 2014; Maples et al., 2013).

In our experiments, we initialised parameters with values extremely close (within 0.0001) to the true simulation parameters to prevent the EM algorithm from becoming trapped in local maxima during optimisation. This approach is effective when true parameters are known, but in practical applications of LOCO, users typically lack this information, as they would not run the simulation mode where parameters are the input. However, a parameter that users can obtain externally and provide to LOCO is the genome-wide ancestry proportion (pi_ip). These proportions represent the overall fraction of an individual's genome that is derived from each ancestral source and are a key component of the transition probabilities in the LOCO model. Instead of asking LOCO to infer these values by the EM optimisation that can fall to local optima, users can compute pi_ip using global ancestry inference software such as ADMIXTURE, fastSTRUCTURE, or STRUCTURE (Alexander et al., 2009; Pritchard et al., 2000; Raj et al., 2014). In the case of ADMIXTURE the output is a matrix where each row corresponds to an individual, and each column corresponds to the proportion of their genome assigned to one of the ancestral populations (Alexander et al., 2009). These global ancestry proportions can be directly interpreted as the pi_ip parameters needed by LOCO. If pi_ip is provided externally by the user, it will reduce the number of parameters LOCO have to infer and reduce the likelihood of converging to a suboptimal local maximum.

Poor initialisation of the parameters can lead to suboptimal inference, preventing the model from converging to the global maximum-likelihood solution. To address this, in future work, we could develop a strategy where we run the EM algorithm multiple times with different random initialisations and select the model that achieves the highest likelihood (Fraley & Raftery, 1998). For HMM, the EM algorithm (specifically Baum-Welch) is highly sensitive to initialisation given the presence of numerous local optima (Bilmes, n.d.). To mitigate these random initialization issues, some methods have been proposed to improve EM initialisation beyond random. One method is Kwedlo's Mahalanobis distance-based approach, originally used in clustering models like Gaussian Mixture Models but also applicable to HMMs (Kwedlo, 2015). The key idea is to choose initial parameter values that are very different from each other so that each hidden state starts off modeling a different part of the data. This is done by measuring how far apart data points are using Mahalanobis distance, which is a way to measure distance that accounts for the shape and spread of the data (unlike regular distance, it considers how variables are correlated). This helps avoid starting two states in the same part of the data, which could lead to poor initialisation to find the global maxima. Another method is Multiple Restart Iterative Partition EM (MRIPEM) (You et al., 2023). Instead of choosing all initial parameters at once, MRIPEM builds the model step by step. It starts with one group and then gradually adds more, each time identifying parts of the data that haven't been well captured yet. After adding each group, it reassigns data points and updates the model parameters. This process helps the model find and separate different patterns in the data early on, making the EM optimisation more stable. In their experiments, You *et al.*. showed that MRIPEM led to more reliable results with fewer differences between runs and better accuracy than standard random starts. These methods can be integrated into LOCO to provide better initialisation of the parameters, helping the model avoid getting stuck in local maxima. At the same time, we can implement the strategy of running multiple initialisations with these alternative methods that will help LOCO explore different regions of the parameter space and select the best-performing solution.

Overall, LOCO's ability to recover diploid ancestry across a wide range of simulated scenarios demonstrates the model's potential for analyzing real-world data where

reference panels may not be available. While challenges remain, these are common to most local ancestry methods and can be addressed through improved initialization strategies.

Authors contributions

Daniel Wegmann (DW) conceived the idea. DW, Madleina Caduff (MC) and C. Sarai Reyes-Avila developed the model. CSRA implemented the methods of the low-depth copy model in collaboration with MC. MC implemented the library of “stattools” that is used by the low-depth copy model. CSRA conducted all simulations and data analyses with the advice and guidance of MC and DW. CSRA led the writing of this manuscript, and DC contributed to the revision of this manuscript.

Newton-Raphson for low-depth copy model

April 13, 2024

Let us denote by $q_{il}^{(X)}(\boldsymbol{\theta})$ the term relevant for case $X \in A, B, C, D, E, F$ of transition probabilities. Here, $\boldsymbol{\theta} = \{\rho^*, \rho_p, \pi_{ip}, q_p, f_{pn}\}$ is the vector of all parameters that are relevant for Newton-Raphson.

Our goal now is to take the first and second derivatives with respect to all these parameters of the function

$$Q(\boldsymbol{\theta}) = \sum_i \sum_l \sum_X \xi_{il}^{(X)} \log\left(q_{il}^{(X)}(\boldsymbol{\theta})\right).$$

Here, $\xi_{il}^{(X)}$ represents the sum over all entries of the ξ_{il} matrix where haplotype (1) and (2) take transition X.

$$\begin{aligned}\xi_{il}^{(X)} &= \sum_{z_{il}^{(1)}} \sum_{z_{il-1}^{(1)}} \sum_{z_{il}^{(2)}} \sum_{z_{il-1}^{(2)}} I\left((z_{il-1}^{(1)} \to z_{il}^{(1)}) \in X\right) \xi_{il}(z_{il}^{(1)}, z_{il}^{(2)} z_{il-1}^{(1)}, z_{il-1}^{(2)}) + \\ &\quad \sum_{z_{il}^{(1)}} \sum_{z_{il-1}^{(1)}} \sum_{z_{il}^{(2)}} \sum_{z_{il-1}^{(2)}} I\left((z_{il-1}^{(2)} \to z_{il}^{(2)}) \in X\right) \xi_{il}(z_{il}^{(1)}, z_{il}^{(2)} z_{il-1}^{(1)}, z_{il-1}^{(2)}).\end{aligned}$$

The following rules apply when taking the derivative for a parameter $\theta \in \boldsymbol{\theta}$:

$$\frac{\partial}{\partial\theta} Q(\boldsymbol{\theta}) = \sum_i \sum_l \sum_X \xi_{il}^{(X)} \frac{\frac{\partial}{\partial\theta} q_{il}^{(X)}}{q_{il}^{(X)}},$$

where $\frac{\partial}{\partial\theta} q_{il}^{(X)}$ corresponds to the first derivative of q_{il} with respect to θ.

$$\frac{\partial^2}{\partial\theta_1 \partial\theta_2} Q(\boldsymbol{\theta}) = \sum_i \sum_l \sum_X \xi_{il}^{(X)} \frac{\frac{\partial^2}{\partial\theta_1\partial\theta_2} q_{il}^{(X)}}{q_{il}^{(X)}} - \frac{\frac{\partial}{\partial\theta_1} q_{il}^{(X)} \frac{\partial}{\partial\theta_2} q_{il}^{(X)}}{{q_{il}^{(X)}}^2},$$

where $\frac{\partial^2}{\partial\theta_1\partial\theta_2} q_{il}^{(X)}$ corresponds to the second derivative of q_{il} with respect to θ_1 and θ_2.

$$\frac{\partial^2}{\partial\theta^2} Q(\boldsymbol{\theta}) = \sum_i \sum_l \sum_X \xi_{il}^{(X)} \left(\frac{\frac{\partial^2}{\partial\theta^2} q_{il}^{(X)}}{q_{il}^{(X)}} - \left(\frac{\frac{\partial}{\partial\theta} q_{il}^{(X)}}{q_{il}^{(X)}} \right)^2 \right),$$

where $\frac{\partial^2}{\partial\theta^2} q_{il}^{(X)}$ corresponds to the second derivative of q_{il} with respect to θ.

1 First derivatives

The first derivatives are:

$$
\begin{aligned}
\frac{\partial}{\partial \rho^*} Q(\boldsymbol{\theta}) \quad = \quad & \sum_i \sum_l \Bigg[\xi_{il}^{(A)} \frac{\bar{R}_l^* \delta_l}{R_l^*} + \dots \\
& + \xi_{il}^{(B)} \frac{\bar{R}_l^* \delta_l}{R_l^*} + \dots \\
& + \xi_{il}^{(C)} \frac{\left(-\bar{R}_l^* \delta_l\right) R_{lp} + \bar{R}_l^* \delta_l \pi_{ip}}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} + \dots \\
& + \xi_{il}^{(D)} \frac{\left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} + \left(-\bar{R}_l^* \delta_l\right) R_{lp} \bar{q}_p f_{pn} + \bar{R}_l^* \delta_l \pi_{ip} \bar{q}_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} + \dots \\
& + \xi_{il}^{(E)} \frac{\left(-\bar{R}_l^* \delta_l\right) R_{lp} + \bar{R}_l^* \delta_l \pi_{ip}}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} + \dots \\
& + \xi_{il}^{(F)} \frac{\left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} + \left(-\bar{R}_l^* \delta_l\right) R_{lp} q_p f_{pn} + \bar{R}_l^* \delta_l \pi_{ip} q_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \rho_p} Q(\boldsymbol{\theta}) \quad = \quad & \sum_i \sum_l \Bigg[\xi_{il}^{(C)} \frac{\bar{R}_l^* \bar{R}_{lp} \delta_l}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} + \dots \\
& + \xi_{il}^{(D)} \frac{\bar{R}_l^* \left(-\bar{R}_{lp} \delta_l\right) + \bar{R}_l^* \bar{R}_{lp} \delta_l \bar{q}_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} + \dots \\
& + \xi_{il}^{(E)} \frac{\bar{R}_l^* \bar{R}_{lp} \delta_l}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} + \dots \\
& + \xi_{il}^{(F)} \frac{\bar{R}_l^* \left(-\bar{R}_{lp} \delta_l\right) + \bar{R}_l^* \bar{R}_{lp} \delta_l q_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \pi_{ip}} Q(\boldsymbol{\theta}) \quad = \quad & \sum_i \sum_l \Bigg[\xi_{il}^{(A)} \frac{1}{\pi_{ip}} + \dots \\
& + \xi_{il}^{(B)} \frac{1}{\pi_{ip}} + \dots \\
& + \xi_{il}^{(C)} \frac{R_l^*}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} + \dots \\
& + \xi_{il}^{(D)} \frac{R_l^* \bar{q}_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} + \dots \\
& + \xi_{il}^{(E)} \frac{R_l^*}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} + \dots \\
& + \xi_{il}^{(F)} \frac{R_l^* q_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial q_p} Q(\boldsymbol{\theta}) \quad = \quad & \sum_i \sum_l \Bigg[\xi_{il}^{(A)} \frac{(-1)}{\bar{q}_p} + \dots \\
& + \xi_{il}^{(B)} \frac{1}{q_p} + \dots \\
& + \xi_{il}^{(C)} \frac{(-1)}{\bar{q}_p} + \dots \\
& + \xi_{il}^{(D)} \frac{\bar{R}_l^* R_{lp} (-1) f_{pn} + R_l^* \pi_{ip} (-1) f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} + \dots \\
& + \xi_{il}^{(E)} \frac{1}{q_p} + \dots \\
& + \xi_{il}^{(F)} \frac{\bar{R}_l^* R_{lp} f_{pn} + R_l^* \pi_{ip} f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial f_{pn}} Q(\boldsymbol{\theta}) &= \sum_i \sum_l \Bigg[\xi_{il}^{(A)} \frac{1}{f_{pn}} + \dots \\
&\quad + \xi_{il}^{(C)} \frac{1}{f_{pn}} + \dots \\
&\quad + \xi_{il}^{(D)} \frac{\bar{R}_l^* R_{lp} \bar{q}_p + R_l^* \pi_{ip} \bar{q}_p}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} + \dots \\
&\quad + \xi_{il}^{(E)} \frac{1}{f_{pn}} + \dots \\
&\quad + \xi_{il}^{(F)} \frac{\bar{R}_l^* R_{lp} q_p + R_l^* \pi_{ip} q_p}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} \Bigg]
\end{aligned}
$$

$$
\frac{\partial}{\partial f_{\tilde{p}n}} Q(\boldsymbol{\theta}) = \sum_i \sum_l \left[\xi_{il}^{(B)} \frac{1}{f_{\tilde{p}n}} \right]
$$

2 Second derivatives

The second derivatives are:

$$
\begin{aligned}
&\frac{\partial^2}{\partial \rho^{*2}} Q(\boldsymbol{\theta}) = \sum_i \sum_l \Bigg[\xi_{il}^{(A)} \left(\frac{\left(-\bar{R}_l^* \delta_l^2\right)}{R_l^*} - \left(\frac{\bar{R}_l^* \delta_l}{R_l^*} \right)^2 \right) + \dots \\
&+ \xi_{il}^{(B)} \left(\frac{\left(-\bar{R}_l^* \delta_l^2\right)}{R_l^*} - \left(\frac{\bar{R}_l^* \delta_l}{R_l^*} \right)^2 \right) + \dots \\
&+ \xi_{il}^{(C)} \left(\frac{\bar{R}_l^* \delta_l^2 R_{lp} + \left(-\bar{R}_l^* \delta_l^2\right) \pi_{ip}}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} - \left(\frac{\left(-\bar{R}_l^* \delta_l\right) R_{lp} + \bar{R}_l^* \delta_l \pi_{ip}}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} \right)^2 \right) + \dots \\
&+ \xi_{il}^{(D)} \left(\frac{\bar{R}_l^* \delta_l^2 \bar{R}_{lp} + \bar{R}_l^* \delta_l^2 R_{lp} \bar{q}_p f_{pn} + \left(-\bar{R}_l^* \delta_l^2\right) \pi_{ip} \bar{q}_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} - \left(\frac{\left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} + \left(-\bar{R}_l^* \delta_l\right) R_{lp} \bar{q}_p f_{pn} + \bar{R}_l^* \delta_l \pi_{ip} \bar{q}_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} \right)^2 \right) + \dots \\
&+ \xi_{il}^{(E)} \left(\frac{\bar{R}_l^* \delta_l^2 R_{lp} + \left(-\bar{R}_l^* \delta_l^2\right) \pi_{ip}}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} - \left(\frac{\left(-\bar{R}_l^* \delta_l\right) R_{lp} + \bar{R}_l^* \delta_l \pi_{ip}}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} \right)^2 \right) + \dots \\
&+ \xi_{il}^{(F)} \left(\frac{\bar{R}_l^* \delta_l^2 \bar{R}_{lp} + \bar{R}_l^* \delta_l^2 R_{lp} q_p f_{pn} + \left(-\bar{R}_l^* \delta_l^2\right) \pi_{ip} q_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} - \left(\frac{\left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} + \left(-\bar{R}_l^* \delta_l\right) R_{lp} q_p f_{pn} + \bar{R}_l^* \delta_l \pi_{ip} q_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} \right)^2 \right) \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial \rho^* \partial \rho_p} Q(\boldsymbol{\theta}) = \sum_i \sum_l \Bigg[\xi_{il}^{(C)} \Bigg(\frac{\left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} \delta_l}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} - \dots \\
&- \frac{\left(\left(-\bar{R}_l^* \delta_l\right) R_{lp} + \bar{R}_l^* \delta_l \pi_{ip}\right) \cdot \left(\bar{R}_l^* \bar{R}_{lp} \delta_l\right)}{\left(\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}\right)^2} \Bigg) + \dots \\
&+ \xi_{il}^{(D)} \Bigg(\frac{\left(-\bar{R}_l^* \delta_l\right) \left(-\bar{R}_{lp} \delta_l\right) + \left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} \delta_l \bar{q}_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} - \dots \\
&- \frac{\left(\left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} + \left(-\bar{R}_l^* \delta_l\right) R_{lp} \bar{q}_p f_{pn} + \bar{R}_l^* \delta_l \pi_{ip} \bar{q}_p f_{pn}\right) \cdot \left(\bar{R}_l^* \left(-\bar{R}_{lp} \delta_l\right) + \bar{R}_l^* \bar{R}_{lp} \delta_l \bar{q}_p f_{pn}\right)}{\left(\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}\right)^2} \Bigg) + \dots \\
&+ \xi_{il}^{(E)} \Bigg(\frac{\left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} \delta_l}{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}} - \dots \\
&- \frac{\left(\left(-\bar{R}_l^* \delta_l\right) R_{lp} + \bar{R}_l^* \delta_l \pi_{ip}\right) \cdot \left(\bar{R}_l^* \bar{R}_{lp} \delta_l\right)}{\left(\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}\right)^2} \Bigg) + \dots \\
&+ \xi_{il}^{(F)} \Bigg(\frac{\left(-\bar{R}_l^* \delta_l\right) \left(-\bar{R}_{lp} \delta_l\right) + \left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} \delta_l q_p f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} - \dots \\
&- \frac{\left(\left(-\bar{R}_l^* \delta_l\right) \bar{R}_{lp} + \left(-\bar{R}_l^* \delta_l\right) R_{lp} q_p f_{pn} + \bar{R}_l^* \delta_l \pi_{ip} q_p f_{pn}\right) \cdot \left(\bar{R}_l^* \left(-\bar{R}_{lp} \delta_l\right) + \bar{R}_l^* \bar{R}_{lp} \delta_l q_p f_{pn}\right)}{\left(\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}\right)^2} \Bigg) \Bigg]
\end{aligned}
$$

$$\frac{\partial^2}{\partial\rho^*\partial\pi_{ip}}Q(\boldsymbol{\theta}) = \sum_i\sum_l\left[\xi_{il}^{(C)}\left(\frac{\bar{R}_l^*\delta_l}{\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}} - \dots\right.\right.$$
$$\left. - \frac{\left(\left(-\bar{R}_l^*\delta_l\right)R_{lp} + \bar{R}_l^*\delta_l\pi_{ip}\right)\cdot\left(R_l^*\right)}{\left(\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}\right)^2}\right) + \dots$$
$$+\xi_{il}^{(D)}\left(\frac{\bar{R}_l^*\delta_l\bar{q}_pf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}} - \dots\right.$$
$$\left. - \frac{\left(\left(-\bar{R}_l^*\delta_l\right)\bar{R}_{lp} + \left(-\bar{R}_l^*\delta_l\right)R_{lp}\bar{q}_pf_{pn} + \bar{R}_l^*\delta_l\pi_{ip}\bar{q}_pf_{pn}\right)\cdot\left(R_l^*\bar{q}_pf_{pn}\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}\right)^2}\right) + \dots$$
$$+\xi_{il}^{(E)}\left(\frac{\bar{R}_l^*\delta_l}{\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}} - \dots\right.$$
$$\left. - \frac{\left(\left(-\bar{R}_l^*\delta_l\right)R_{lp} + \bar{R}_l^*\delta_l\pi_{ip}\right)\cdot\left(R_l^*\right)}{\left(\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}\right)^2}\right) + \dots$$
$$+\xi_{il}^{(F)}\left(\frac{\bar{R}_l^*\delta_lq_pf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}} - \dots\right.$$
$$\left.\left. - \frac{\left(\left(-\bar{R}_l^*\delta_l\right)\bar{R}_{lp} + \left(-\bar{R}_l^*\delta_l\right)R_{lp}q_pf_{pn} + \bar{R}_l^*\delta_l\pi_{ip}q_pf_{pn}\right)\cdot\left(R_l^*q_pf_{pn}\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}\right)^2}\right)\right]$$

$$\frac{\partial^2}{\partial\rho^*\partial q_p}Q(\boldsymbol{\theta}) = \sum_i\sum_l\left[\xi_{il}^{(D)}\left(\frac{\left(-\bar{R}_l^*\delta_l\right)R_{lp}\left(-1\right)f_{pn} + \bar{R}_l^*\delta_l\pi_{ip}\left(-1\right)f_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}} - \dots\right.\right.$$
$$\left. - \frac{\left(\left(-\bar{R}_l^*\delta_l\right)\bar{R}_{lp} + \left(-\bar{R}_l^*\delta_l\right)R_{lp}\bar{q}_pf_{pn} + \bar{R}_l^*\delta_l\pi_{ip}\bar{q}_pf_{pn}\right)\cdot\left(\bar{R}_l^*R_{lp}\left(-1\right)f_{pn} + R_l^*\pi_{ip}\left(-1\right)f_{pn}\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}\right)^2}\right) + \dots$$
$$+\xi_{il}^{(F)}\left(\frac{\left(-\bar{R}_l^*\delta_l\right)R_{lp}f_{pn} + \bar{R}_l^*\delta_l\pi_{ip}f_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}} - \dots\right.$$
$$\left.\left. - \frac{\left(\left(-\bar{R}_l^*\delta_l\right)\bar{R}_{lp} + \left(-\bar{R}_l^*\delta_l\right)R_{lp}q_pf_{pn} + \bar{R}_l^*\delta_l\pi_{ip}q_pf_{pn}\right)\cdot\left(\bar{R}_l^*R_{lp}f_{pn} + R_l^*\pi_{ip}f_{pn}\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}\right)^2}\right)\right]$$

$$\frac{\partial^2}{\partial\rho^*\partial f_{pn}}Q(\boldsymbol{\theta}) = \sum_i\sum_l\left[\xi_{il}^{(D)}\left(\frac{\left(-\bar{R}_l^*\delta_l\right)R_{lp}\bar{q}_p + \bar{R}_l^*\delta_l\pi_{ip}\bar{q}_p}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}} - \dots\right.\right.$$
$$\left. - \frac{\left(\left(-\bar{R}_l^*\delta_l\right)\bar{R}_{lp} + \left(-\bar{R}_l^*\delta_l\right)R_{lp}\bar{q}_pf_{pn} + \bar{R}_l^*\delta_l\pi_{ip}\bar{q}_pf_{pn}\right)\cdot\left(\bar{R}_l^*R_{lp}\bar{q}_p + R_l^*\pi_{ip}\bar{q}_p\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}\right)^2}\right) + \dots$$
$$+\xi_{il}^{(F)}\left(\frac{\left(-\bar{R}_l^*\delta_l\right)R_{lp}q_p + \bar{R}_l^*\delta_l\pi_{ip}q_p}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}} - \dots\right.$$
$$\left.\left. - \frac{\left(\left(-\bar{R}_l^*\delta_l\right)\bar{R}_{lp} + \left(-\bar{R}_l^*\delta_l\right)R_{lp}q_pf_{pn} + \bar{R}_l^*\delta_l\pi_{ip}q_pf_{pn}\right)\cdot\left(\bar{R}_l^*R_{lp}q_p + R_l^*\pi_{ip}q_p\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}\right)^2}\right)\right]$$

$$\frac{\partial^2}{\partial{\rho_p}^2}Q(\boldsymbol{\theta}) = \sum_i\sum_l\left[\xi_{il}^{(C)}\left(\frac{\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l^2\right)}{\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}} - \left(\frac{\bar{R}_l^*\bar{R}_{lp}\delta_l}{\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}}\right)^2\right) + \dots\right.$$
$$+\xi_{il}^{(D)}\left(\frac{\bar{R}_l^*\bar{R}_{lp}\delta_l^2 + \bar{R}_l^*\left(-\bar{R}_{lp}\delta_l^2\right)\bar{q}_pf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}} - \left(\frac{\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l\right) + \bar{R}_l^*\bar{R}_{lp}\delta_l\bar{q}_pf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}}\right)^2\right) + \dots$$
$$+\xi_{il}^{(E)}\left(\frac{\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l^2\right)}{\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}} - \left(\frac{\bar{R}_l^*\bar{R}_{lp}\delta_l}{\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}}\right)^2\right) + \dots$$
$$\left.+\xi_{il}^{(F)}\left(\frac{\bar{R}_l^*\bar{R}_{lp}\delta_l^2 + \bar{R}_l^*\left(-\bar{R}_{lp}\delta_l^2\right)q_pf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}} - \left(\frac{\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l\right) + \bar{R}_l^*\bar{R}_{lp}\delta_lq_pf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}}\right)^2\right)\right]$$

$$\frac{\partial^2}{\partial\rho_p\partial\pi_{ip}}Q(\boldsymbol{\theta}) = \sum_i\sum_l\left[\xi_{il}^{(C)}\frac{-\left(\bar{R}_l^*\bar{R}_{lp}\delta_l\right)\cdot\left(R_l^*\right)}{\left(\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}\right)^2}+\dots\right.$$
$$+\xi_{il}^{(D)}\frac{-\left(\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l\right)+\bar{R}_l^*\bar{R}_{lp}\delta_l\bar{q}_pf_{pn}\right)\cdot\left(R_l^*\bar{q}_pf_{pn}\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}\right)^2}+\dots$$
$$+\xi_{il}^{(E)}\frac{-\left(\bar{R}_l^*\bar{R}_{lp}\delta_l\right)\cdot\left(R_l^*\right)}{\left(\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}\right)^2}+\dots$$
$$\left.+\xi_{il}^{(F)}\frac{-\left(\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l\right)+\bar{R}_l^*\bar{R}_{lp}\delta_lq_pf_{pn}\right)\cdot\left(R_l^*q_pf_{pn}\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}\right)^2}\right]$$

$$\frac{\partial^2}{\partial\rho_p\partial q_p}Q(\boldsymbol{\theta}) = \sum_i\sum_l\left[\xi_{il}^{(D)}\left(\frac{\bar{R}_l^*\bar{R}_{lp}\delta_l\left(-1\right)f_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}}-\dots\right.\right.$$
$$\left.-\frac{\left(\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l\right)+\bar{R}_l^*\bar{R}_{lp}\delta_l\bar{q}_pf_{pn}\right)\cdot\left(\bar{R}_l^*R_{lp}\left(-1\right)f_{pn}+R_l^*\pi_{ip}\left(-1\right)f_{pn}\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}\right)^2}\right)+\dots$$
$$+\xi_{il}^{(F)}\left(\frac{\bar{R}_l^*\bar{R}_{lp}\delta_lf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}}-\dots\right.$$
$$\left.\left.-\frac{\left(\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l\right)+\bar{R}_l^*\bar{R}_{lp}\delta_lq_pf_{pn}\right)\cdot\left(\bar{R}_l^*R_{lp}f_{pn}+R_l^*\pi_{ip}f_{pn}\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}\right)^2}\right)\right]$$

$$\frac{\partial^2}{\partial\rho_p\partial f_{pn}}Q(\boldsymbol{\theta}) = \sum_i\sum_l\left[\xi_{il}^{(D)}\left(\frac{\bar{R}_l^*\bar{R}_{lp}\delta_l\bar{q}_p}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}}-\dots\right.\right.$$
$$\left.-\frac{\left(\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l\right)+\bar{R}_l^*\bar{R}_{lp}\delta_l\bar{q}_pf_{pn}\right)\cdot\left(\bar{R}_l^*R_{lp}\bar{q}_p+R_l^*\pi_{ip}\bar{q}_p\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}\right)^2}\right)+\dots$$
$$+\xi_{il}^{(F)}\left(\frac{\bar{R}_l^*\bar{R}_{lp}\delta_lq_p}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}}-\dots\right.$$
$$\left.\left.-\frac{\left(\bar{R}_l^*\left(-\bar{R}_{lp}\delta_l\right)+\bar{R}_l^*\bar{R}_{lp}\delta_lq_pf_{pn}\right)\cdot\left(\bar{R}_l^*R_{lp}q_p+R_l^*\pi_{ip}q_p\right)}{\left(\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}\right)^2}\right)\right]$$

$$\frac{\partial^2}{\partial{\pi_{ip}}^2}Q(\boldsymbol{\theta}) = \sum_i\sum_l\left[\xi_{il}^{(A)}\left(-\left(\frac{1}{\pi_{ip}}\right)^2\right)+\dots\right.$$
$$+\xi_{il}^{(B)}\left(-\left(\frac{1}{\pi_{ip}}\right)^2\right)+\dots$$
$$+\xi_{il}^{(C)}\left(-\left(\frac{R_l^*}{\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}}\right)^2\right)+\dots$$
$$+\xi_{il}^{(D)}\left(-\left(\frac{R_l^*\bar{q}_pf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}\bar{q}_pf_{pn}+R_l^*\pi_{ip}\bar{q}_pf_{pn}}\right)^2\right)+\dots$$
$$+\xi_{il}^{(E)}\left(-\left(\frac{R_l^*}{\bar{R}_l^*R_{lp}+R_l^*\pi_{ip}}\right)^2\right)+\dots$$
$$\left.+\xi_{il}^{(F)}\left(-\left(\frac{R_l^*q_pf_{pn}}{\bar{R}_l^*\bar{R}_{lp}+\bar{R}_l^*R_{lp}q_pf_{pn}+R_l^*\pi_{ip}q_pf_{pn}}\right)^2\right)\right]$$

$$\frac{\partial^2}{\partial \pi_{ip} \partial q_p} Q(\boldsymbol{\theta}) = \sum_i \sum_l \left[\xi_{il}^{(D)} \left(\frac{R_l^* \left(-1\right) f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} - \dots \right.\right.$$
$$\left. - \frac{\left(R_l^* \bar{q}_p f_{pn}\right) \cdot \left(\bar{R}_l^* R_{lp} \left(-1\right) f_{pn} + R_l^* \pi_{ip} \left(-1\right) f_{pn}\right)}{\left(\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}\right)^2} \right) + \dots$$
$$+ \xi_{il}^{(F)} \left(\frac{R_l^* f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} - \dots \right.$$
$$\left.\left. - \frac{\left(R_l^* q_p f_{pn}\right) \cdot \left(\bar{R}_l^* R_{lp} f_{pn} + R_l^* \pi_{ip} f_{pn}\right)}{\left(\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}\right)^2} \right) \right]$$

$$\frac{\partial^2}{\partial \pi_{ip} \partial f_{pn}} Q(\boldsymbol{\theta}) = \sum_i \sum_l \left[\xi_{il}^{(D)} \left(\frac{R_l^* \bar{q}_p}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} - \dots \right.\right.$$
$$\left. - \frac{\left(R_l^* \bar{q}_p f_{pn}\right) \cdot \left(\bar{R}_l^* R_{lp} \bar{q}_p + R_l^* \pi_{ip} \bar{q}_p\right)}{\left(\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}\right)^2} \right) + \dots$$
$$+ \xi_{il}^{(F)} \left(\frac{R_l^* q_p}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} - \dots \right.$$
$$\left.\left. - \frac{\left(R_l^* q_p f_{pn}\right) \cdot \left(\bar{R}_l^* R_{lp} q_p + R_l^* \pi_{ip} q_p\right)}{\left(\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}\right)^2} \right) \right]$$

$$\begin{aligned}
\frac{\partial^2}{\partial {q_p}^2} Q(\boldsymbol{\theta}) = \sum_i \sum_l \Bigg[& \xi_{il}^{(A)} \left(- \left(\frac{\left(-1\right)}{\bar{q}_p} \right)^2 \right) + \dots \\
& + \xi_{il}^{(B)} \left(- \left(\frac{1}{q_p} \right)^2 \right) + \dots \\
& + \xi_{il}^{(C)} \left(- \left(\frac{1}{\bar{q}_p} \right)^2 \right) + \dots \\
& + \xi_{il}^{(D)} \left(- \left(\frac{\bar{R}_l^* R_{lp} \left(-1\right) f_{pn} + R_l^* \pi_{ip} \left(-1\right) f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} \right)^2 \right) + \dots \\
& + \xi_{il}^{(E)} \left(- \left(\frac{1}{q_p} \right)^2 \right) + \dots \\
& + \xi_{il}^{(F)} \left(- \left(\frac{\bar{R}_l^* R_{lp} f_{pn} + R_l^* \pi_{ip} f_{pn}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} \right)^2 \right) \Bigg]
\end{aligned}$$

$$\frac{\partial^2}{\partial q_p \partial f_{pn}} Q(\boldsymbol{\theta}) = \sum_i \sum_l \left[\xi_{il}^{(D)} \left(\frac{\bar{R}_l^* R_{lp} \left(-1\right) + R_l^* \pi_{ip} \left(-1\right)}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} - \dots \right.\right.$$
$$\left. - \frac{\left(\bar{R}_l^* R_{lp} \left(-1\right) f_{pn} + R_l^* \pi_{ip} \left(-1\right) f_{pn}\right) \cdot \left(\bar{R}_l^* R_{lp} \bar{q}_p + R_l^* \pi_{ip} \bar{q}_p\right)}{\left(\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}\right)^2} \right) + \dots$$
$$+ \xi_{il}^{(F)} \left(\frac{\bar{R}_l^* R_{lp} + R_l^* \pi_{ip}}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} - \dots \right.$$
$$\left.\left. - \frac{\left(\bar{R}_l^* R_{lp} f_{pn} + R_l^* \pi_{ip} f_{pn}\right) \cdot \left(\bar{R}_l^* R_{lp} q_p + R_l^* \pi_{ip} q_p\right)}{\left(\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}\right)^2} \right) \right]$$

$$\frac{\partial^2}{\partial {f_{pn}}^2} Q(\boldsymbol{\theta}) = \sum_i \sum_l \left[\xi_{il}^{(A)} \left(- \left(\frac{1}{f_{pn}} \right)^2 \right) + \dots \right.$$
$$+ \xi_{il}^{(C)} \left(- \left(\frac{1}{f_{pn}} \right)^2 \right) + \dots$$
$$+ \xi_{il}^{(D)} \left(- \left(\frac{\bar{R}_l^* R_{lp} \bar{q}_p + R_l^* \pi_{ip} \bar{q}_p}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} \bar{q}_p f_{pn} + R_l^* \pi_{ip} \bar{q}_p f_{pn}} \right)^2 \right) + \dots$$
$$+ \xi_{il}^{(E)} \left(- \left(\frac{1}{f_{pn}} \right)^2 \right) + \dots$$
$$\left. + \xi_{il}^{(F)} \left(- \left(\frac{\bar{R}_l^* R_{lp} q_p + R_l^* \pi_{ip} q_p}{\bar{R}_l^* \bar{R}_{lp} + \bar{R}_l^* R_{lp} q_p f_{pn} + R_l^* \pi_{ip} q_p f_{pn}} \right)^2 \right) \right]$$

$$\frac{\partial^2}{\partial {f_{\tilde{p}n}}^2} Q(\boldsymbol{\theta}) \quad = \quad \sum_i \sum_l \left[\xi_{il}^{(B)} \left(- \left(\frac{1}{f_{pn}} \right)^2 \right) \right]$$

CHAPTER 3

Machine Learning for Detection of Next-Generation Genetically Modified Organisms Through Ancestry Inference

C. Sarai Reyes-Avila [1], Madleina Caduff [2], Daniel Wegmann[2] , Daniel Croll[1],

[1]Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, CH-2000, Neuchâtel, Switzerland

[2] Statistical and Computational Biology, Department of Biology, University of Fribourg, CH-1700, Fribourg, Switzerland

ABSTRACT: The production and trade of genetically modified (GM) crops are regulated by governments globally with a high degree of variation in rules. It affects the planting, marketing, labelling, and trading of GMs. To ensure compliance, accurate detection and screening methods for GMs are necessary. Existing molecular methods efficiently screen traditional GMs. However, the advance of New Genomic Techniques (NGTs), including CRISPR/Cas9, introduce only minimal modifications, called second-generation GMOs. Second-generation GMOs are challenging to screen for if the modifications are not previously known. Here, we explore whether such screenings are still feasible if the species context is taken into account. We apply an ancestry inference method for screening second-generation GMO-like modifications as a proof-of-principle approach with an algorithm called low-depth copy model (LOCO). LOCO does not require a reference panel and uses genotype likelihoods as input suitable for low-input sequencing data. We focused on a diverse sample of unmodified rice cultivars and artificially created second-generation GMO-like modifications by swapping chromosomal segments. We discuss the detection limits considering the scope of modifications, the genomic locations and the sequencing coverage. Future applications should include real-world second-generation GMOs modifications and assess the suitability to implement the procedure in second-generation GMOs screening efforts.

INTRODUCTION

Genetically modified organisms (GMOs) are developed for agricultural and industrial applications. Ninety-nine percent of GM crops that are commercially available have characteristics including resistance to abiotic stress, herbicides, and insects. (Rozas et al., 2022). Herbicide tolerance, for instance, is provided by genetically modifying soybeans to express a gene that codes for 5-enolpyruvylshikimate-3-phosphate synthase, which confers resistance to herbicides based on glyphosate. *Agrobacterium sp.* strain CP4, which is naturally resistant to glyphosate, is the source of the gene. The gene was introduced into the plant genome by transforming the plant using *Agrobacterium tumefaciens.* (Arias et al., 2021). As an example of insect resistance, corn was modified to express *Bacillus thuringiensis's* Cry1Ab protein to defend against pests such as the European corn borer.(Priestley & Brownbridge, 2009). Canola is another example; it was modified to have a different fatty acid makeup, which increased its industrial uses. (Neff et al., 1994). Apart from these traits, plants have also undergone modifications to enhance their nutritional value. For instance, in areas where rice is the main food supply for humans, the "golden rice" was created to manufacture provitamin A in order to alleviate vitamin A deficits, similar to the "golden banana”

production. (Ye et al., 2000) (Paul et al., 2017). However, all of these GMOs were developed by traditional genetic engineering methods that insert genetic constructs, including promoters, terminators, and coding sequences. These GMOs are called first-generation GMOs. The origin of the inserted sequence is typical of heterologous origin.

The second generation of GMOs was produced via so-called New Genomic Techniques (NGT), including CRISPR/Cas9 (Bohle et al., 2024). Second-generation GMOs have targeted edits; NGT allows editing a small number of base-pairs or excise sequences rather than introducing permanently exogeneous DNA. Typical modifications achieved through CRISPR/Cas9 in agriculture include single-gene knockouts, precise nucleotide substitutions, and small deletions (Arora & Narula, 2017). CRISPR/Cas9 editing of hexaploid bread wheat at all three homoeo-alleles of the MLO gene is one agricultural application providing broad-spectrum resistance to powdery mildew. (Y. Wang et al., 2014). In tomatoes, for increasing the soluble sugar content, CRISPR/Cas9 was used to knock out the genes SlINVINH1 and SlVPE5(B. Wang et al., 2021). Furthermore, in tomatoes CRISPR/Cas9 was used to edit specific carotenoid biosynthesis genes to enhance the lycopene content and red colour of fruits (Tiwari et al., 2023). In bananas, to prevent the activation of infectious viral particles CRISPR/Cas9 was used to inactivate the endogenous banana streak virus (eBSV) (Tripathi et al., 2019). The development of GM rice (*Oryza sativa*) varieties highlights the adoption of CRISPR/Cas9, representing 12-33% of global gene-editing research compared to other crops (Chen et al., 2024). Many genes have been edited in rice to enhance various traits: Pi21 was knocked out to confer resistance to rice blast, OsSWEET14 was edited to improve resistance against bacterial blight caused by *Xanthomonas oryzae*, OsSPL10 was knocked out to enhance resistance to the brown planthopper, OsRR22 was modified to enhance salt tolerance significantly, GS9 was edited to increase grain length, and the Wx promoter region was edited to influence the texture and taste of cooked rice (F. Wang et al., 2016, Zeng et al., 2020, Lu et al., 2018, Zhang et al., 2019, Zhao et al., 2018, Zeng et al., 2020).

Despite numerous potential applications of second-generation GMOs in agriculture, the regulatory landscape has not evolved substantiarily in Europe (Koller & Cieslak, 2023).At first, all gene-edited organisms were categorised as GMOs by the European Union (EU), which applied the same rules. (Schmidt et al., 2020). The primary law governing the release of genetically modified organisms into the environment in the European Union is Directive 2001/18/EC. This directive aimed to control first-generation genetically modified organisms (GMOs) created by previous genetic engineering methods. Developments in NGTs produced second-generation GMOs, which are genetic alterations that might be identical to those found in nature. These developments raised questions about whether the current legal system is appropriate for regulating genetically modified organisms of the second generation. (New Genomic Techniques, 2024). Directive 2001/18 is being revised to improve the legal framework for gene-edited crops. (New Genomic Techniques, 2024). While genetic engineering continues advancing, the legal framework continues adapting. The detection of first-generation GMOs is well-established, relying on techniques that identify known genetic constructs and heterologous DNA. Quantitative polymerase chain reactions (qPCR) is widely used due to their sensitivity and specificity to detect genetic modifications (Broeders et al., 2012; Marmiroli et al., 2008). Such PCR-based techniques are also the primary current method to detect second-generation GMOs(Chhalliyil et al., 2020). For the first commercially available genome-edited plant, canola, single-nucleotide changes have been quantified using a Real-Time Quantitive PCR (qPCR) (Chhalliyil et al., 2020). However, PCR methods share the limitation of requiring prior knowledge of the genetic modification to design specific primers, hence limiting the identification of unknown genome edits (Broeders et al., 2012). Also, the use of qPCR is problematic for detecting edits introduced by NGT because they involve single nucleotide modifications that are difficult to detect by standard qPCR assays (Aubry et al., 2021). Next-generation sequencing (NGS) helps to detect first-generation GMOs, determine insertion locations, copy number variations, and possible off-target mutations. For example, a multiplex amplification combined with NGS sequencing detected numerous transgenic constructs of first-generation GMOs (Reyes-Avila et al., 2023). Another example is Fraiture *et al*., combining DNA walking to amplify unknown

transgene flanking regions with high-throughput sequencing. Their method is capable of distinguishing transgene configurations and flanking regions, enabling the detection of unauthorised GMOs (Fraiture et al., 2017). The high resolution of NGS can be useful for the detection of second-generation GMOs (Grohmann et al., 2019). Whole-genome sequencing (WGS) can help to detect modifications without prior knowledge. (Grohmann et al., 2019). The procedure involves matching the sample's sequencing reads to a reference genome sequence to identify possible genome modifications. However, it can be challenging to tell the difference between the result of genome editing and spontaneous mutations if the changes come from a closely related species. Existing PCR-based and NGS-based methods often rely on having prior knowledge of specific introduced constructs or clear divergence from a known reference, which is not always the case for second-generation GMOs that may have minimal or naturally occurring variations. Moreover, when the introduced DNA originates from closely related species or when only single-nucleotide edits are introduced, these changes may be indistinguishable from the genetic background variation, thus limiting the sensitivity of traditional detection approaches. Given this, there is a strong demand to improve detection procedures and develop new statistical algorithms addressing this challenge.

Ancestry methods help to assign the genetic contributions of ancestral populations to admixed individuals. Reference panels are a collection of genotyping data from individuals with well-characterized and distinct ancestral backgrounds (Salter-Townshend & Myers, 2019). Regardless of the progress of ancestry inference methods, reference panels are needed; this represents a challenge for no model organisms where reference panels are not always available (all the references). The Low-depth Copy Model (LOCO) algorithm determines ancestry by dynamically reconstructing the reference panels, hence abolishing the need for *a priori* reference panels. LOCO was also proven to detect introgression events. A reference-panel-free approach is valuable for second-generation GMOs, where known transgenes or well-characterized markers might be absent, and reference panels may not be readily available. This approach has the potential to detect genome modifications in second-generation GMOs. LOCO can detect subtle genomic changes, inferring recombination events based on learning recombination parameters from the genotype

likelihood data. Such subtle genomic changes are often found in second-generation GMOs. By focusing on variation patterns rather than searching for a single, known transgene, LOCO can potentially detect edits shared with closely related lineages, offering a powerful complementary method to existing GMO detection pipelines. Furthermore, LOCO may be able to detect genome edits originating from closely related lineages within the same species. An additional advantage of LOCO is the ability to use genotype likelihoods as typically obtained from low-quality sequencing data. Real-world testing for second-generation GMOs content may have to be performed on contaminated or low DNA input samples. Hence, the genotype likelihoods can circumvent the challenge of hard-calling genotypes.

In this study, we aim to evaluate the performance of the LOCO model to detect subtle genetic modification representative of second-generation GMOs. We selected a dataset from rice that contains closely related individuals to simulate the genome edits that come from the introgression of genes from closely related species. We tested LOCO under 10x coverage, simulating levels of poor-quality data. To recreate second-generation GMOs modifications and have control over the genetic modifications, we copied genome segments from a donor individual to other genotypes.

METHODS

Sample selection

The selected dataset comes from the “3000 Rice Genome Project” (The 3, 2014). From this dataset, we subsampled 55 individuals from the "Japonica" variety group, representing two distinct populations: 45 samples of Japonica individuals from Indonesia and 10 samples of Japonica individuals from Japan.

Data processing

The raw sequencing data was obtained from the European Nucleotide Archive (ENA) at EMBL-EBI using SRA Toolkit and the fastq-dump tool. The *Oryza indica* genome

consists of 427 million base pairs across 12 chromosomes, and sequencing reads were 83 base pairs long. We divided the genome length by the sequencing read length to obtain the total number of reads required to achieve 10x coverage (*n* = 51,445,780 reads). We then used the fastq-dump tool to extract paired-end reads, limiting output to the required coverage levels. We downloaded the reference genome for *O. sativa* japonica (IRGSP-1.0) from Ensembl and indexed it using BWA (H. Li, 2013). Reads were aligned to Chromosome 1 using BWA-MEM (H. Li, 2013). The resulting SAM files were converted to BAM format, sorted, and indexed using SAMtools (H. Li et al., 2009). Duplicate reads were identified and marked using GATK MarkDuplicates (McKenna et al., 2010). To call genotype likelihoods and produce BEAGLE format output, we utilised ANGSD (Korneliussen et al., 2014). Genotype likelihood estimation was conducted using the following parameters: GL 1, doGlf 2, doMajorMinor 1, doMaf 1, SNP_pval 1e-6, minMaf 0.05, minQ 20, and minMapQ 30.

Simulation of second-generation GMOs-like

We processed the BEAGLE file generated during data processing to simulate second-generation GMO-like modifications. The analysis was restricted to loci between positions 3,901,986 and 8,243,925 on chromosome 1, corresponding to a subset of 4000 SNPs. We selected a subset of Japonica individuals from Japan: Individual 50, 51, 52, 53, 54, corresponding to ERS468422, ERS468423, ERS468424, ERS468425, ERS468426 respectively. Their genotype likelihoods were duplicated to create five additional individuals: Individual 55, 56, 57, 58, and 59. Artificial introgression was introduced by modifying the SNPs calleds of the duplicated Japonica individuals with donor segments from Individual 0, originating from Indonesia. We simulated two introgression events: one introgression of 100 SNP from SNP index 500 to 600 (as ordered on the chromosome) and a smaller introgression of 50 SNPs at SNP indices 500 to 550.

Implementation of LOCO for detecting second-generation GMOs-like

The low-depth copy model (LOCO) was run to detect introgression-like events in the Japonica subset that included the second-generation GMOs-like. The initialised parameters for running LOCO were: number of haplotypes per population 2, 2; maximum number of iterations 100; and minimum log-likelihood convergence threshold of 0.0001. The output of LOCO was a file with the state probabilities over all combined states for each individual and each locus.

Extraction of the population of origin

We retrieved ancestry calls from the posterior distribution over all combined hidden states in order to obtain the population of origin. A diploid configuration of two haplotypes, denoted as (p1,p_tilde1,n1)×(p2,p_tilde2,n2), is represented by each state. In order to determine the true ancestral source, we marginalized the posterior distribution by focusing on the p1 and p2 components of each state, which represent the real population from which haplotypes 1 and 2 were inherited. To determine the diploid ancestry calls, we listed all possible single-haplotype states based on the number of populations and haplotypes per population. We next constructed a population-by-population matrix using the posterior probabilities for each individual and each locus. Each entry (i,j) in this matrix indicates the overall likelihood that haplotypes 1 and 2 originated from populations i and j, respectively. This gave us the diploid ancestry call, for example, for two populations: P1/P1, P1/P2, or P2/P2 for every individual at each locus.

RESULTS

Detection of a 100-loci Introgression in Second-generation GMO-like Genomes

To evaluate the ability of the LOCO to detect artificially introduced genome modifications mimicking second-generation GMOs, we simulated a longer introgression event consisting of 100 SNP loci (SNP index 500–600). This introgression was introduced into five Japonica individuals (Individual 55 to Individual 59) by copying a chromosomal segment from an Indonesian donor Individual 0. After running LOCO with this data the ancestry of each individual at each locus was represented by one of three possible diploid ancestry states: P1/P1 (homozygous), P1/P2(heterozygous), or P2/P2 (homozygous), which correspond to the combinations of haplotypes inherited from two ancestral populations as we inputed to LOCO (Figure 1). To assess LOCO's performance, we compared the inferred diploid ancestry of each second-generation GMO-like individual against the ancestry of the donor. Individual 56 showed the highest match, with 86 out of 100 loci correctly inferred. This was followed by Individual 55 with 65 matches. In contrast, Individuals 57 to 59 have only 10 matching loci each.

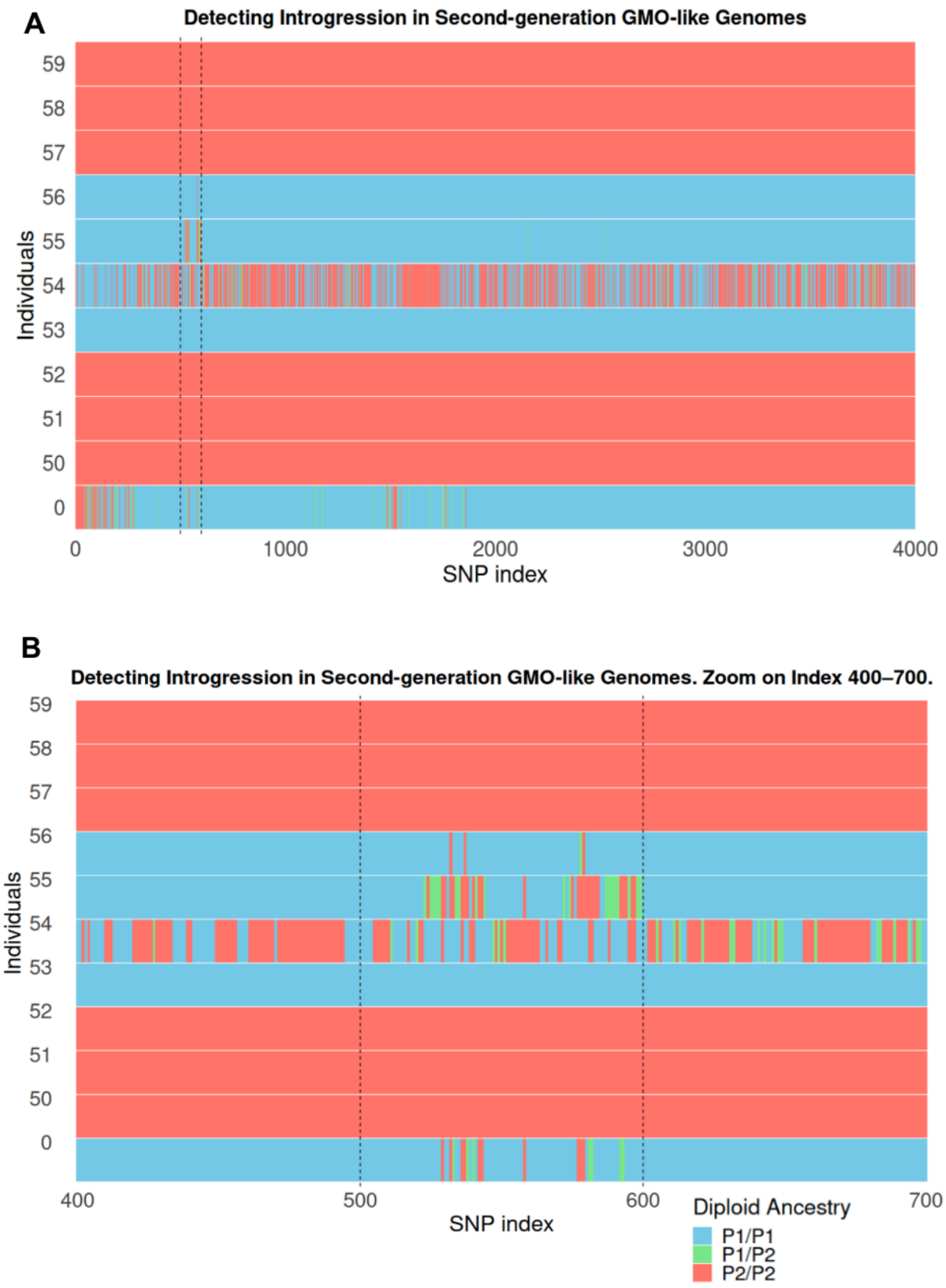
A
Detecting Introgression in Second-generation GMO-like Genomes
Individuals
59
58
57
56
55
54
53
52
51
50
0
0
1000
2000
3000
4000
SNP index
B
Detecting Introgression in Second-generation GMO-like Genomes. Zoom on Index 400–700.
Individuals
59
58
57
56
55
54
53
52
51
50
0
400
500
600
700
SNP index
Diploid Ancestry
P1/P1
P1/P2
P2/P2

Figure 1. Detecting Introgression in Second-generation GMO-like Genomes. (A) Each row represents an individual, and each column represents an SNP lindex (total of 4,000 SNPs). The color indicates the inferred diploid ancestry state at each locus: P1/P1, P1/P2, or P2/P2. Individuals 55–59 were modified to simulate a 100-loci introgression from donor Individual 0 by copying a chromosomal segment from SNP index 500 to 600. Individuals 50–54 are unmodified Japonica controls. The vertical dashed lines mark the introgressed region. (B) Close-up of the region SNP indices 400 to 700 from Figure (A), focusing on the simulated introgression.

Detection of a 50-loci Introgression in Second-generation GMO-like Genomes

To further evaluate the sensitivity of LOCO to detect smaller-scale genome edits, we simulated a shorter introgression event consisting of 50 SNP loci (SNP index 500–550). As in the previous experiment, the introgressed segment was copied from an Indonesian donor (Individual 0) into five Japonica individuals (Individuals 55 to 59). After running LOCO on this dataset, diploid ancestry at each locus was inferred as one of three possible states: P1/P1, P1/P2, or P2/P2, based on the two ancestral populations input in the model (Figure 3). The inferred ancestry patterns for individual 56 have the highest match with the donor, with 41 out of 50 loci. Individual 55 followed closely, with 37 matches. In contrast, Individuals 57 to 59 showed only 6 matching loci each.


A
Detecting Introgression in Second-generation GMO-like Genomes
Individuals
59
58
57
56
55
54
53
52
51
50
0
0
1000
2000
3000
4000
SNP index
B
Detecting Introgression in Second-generation GMO-like Genomes. Zoom on Index 400–600.
Individuals
59
58
57
56
55
54
53
52
51
50
0
400
450
500
550
600
SNP index
Diploid Ancestry
P1/P1
P1/P2
P2/P2

DISCUSSION

This study used the low-depth copy model (LOCO) to detect introgressed segments in second-generation GMO-like genomes. Our primary goal was to determine whether our LOCO approach could effectively recover a genomic segment copied from an Indonesian donor (Individual 0) into Japonica individuals (Individuals 50–59). However, in contrast to our previous simulations from the LOCO manuscript, the ancestry inference in this study was inconsistent. Individuals 0 to 45 are coming from Indonesia and Individuals 46 to 54 are coming from Japan. The model should have recognised these as two separate ancestral populations, one representing the Indonesian group and the other the Japanese group. However, LOCO misassigned the ancestry between them instead of detecting a clear and distinct ancestry assignment for these two populations. Individuals from Indonesia and Japan were indifferently assigned ancestry P1/P1 and P2/2. We only utilised 4000 SNPs from chromosome 1, representing a fraction of the whole genome. This subset of loci may not carry sufficient population-specific signals to distinguish between geographically distinct populations like Indonesians and Japanese. The limited data may constrain the model's ability to differentiate the ancestry sources during parameter estimation. Studies from population genetics support the concern that using a limited number of SNPs from a small genomic region can lead to a misassignment of ancestry (Mizuno et al., 2021; Turakulov & Easteal, 2003). To identify broad population structure, Turakulov and Easteal showed that a minimum of 65 randomly chosen SNP loci are needed (Turakulov & Easteal, 2003). Increasing the number to more than 100 SNP increases the likelihood of correctly assigning individuals to their corresponding population groups, even with simple clustering techniques. Their work emphasised that the power to detect population structure depends heavily on the number and informativeness of the loci analysed. Moreover, when distinguishing between closely related populations, like Japanese versus other East Asians, more markers are required. Mizuno *et al.* showed that distinguishing Japanese individuals from other East Asian populations (e.g., Chinese Dai, Han Chinese, Kinh Vietnamese) required approximately 3000 randomly selected SNPs across the genome to achieve reliable separation using principal

component analysis (Mizuno et al., 2021). Their results imply that subtle genetic differences between geographically proximate populations may be undetectable without a sufficient number of markers, even when using genome-wide data. In addition to the limited genomic region of 4000 SNP, LOCO may converge to suboptimal local maxima in the EM parameter space if the initialisation is not robust. In Chapter 2, we set the initialisation very close to the true parameters (perturbed by 0.0001). However, the true parameter values are unknown in practical applications, as in Chapter 3. The EM algorithm may settle on a local maximum without a robust strategy, resulting in inaccurate ancestry estimates. One solution is to perform multiple random initialisations, running LOCO with its EM algorithm several times with different starting parameters and then selecting the run with the highest log likelihood as the best solution. Wu and Zhou studied a situation with two overlapping groups (Wu & Zhou, 2019). They found that by starting the EM algorithm with random values and using a large enough dataset, the algorithm will likely get very close to the best solution after about the square root of the number of data points in iterations. Similarly, Daskalakis *et al.* demonstrated that as few as ten EM iterations can suffice for mixtures of two Gaussians when multiple random initialisations are employed, thereby supporting the strategy of using multiple random restarts to improve convergence to the global maximum in EM-based models (Daskalakis et al., 2017). These studies support the idea that using multiple random starting points in the EM algorithm can help it find the best possible solutions, even in complex situations.

Because the ancestry inference itself is inaccurately estimated, the detection of the introgressed segment is ambiguous. For instance, an individual that should be predominantly a PX/PX coming from Indonesian (Individual 0) but becomes misassigned to a PY/PY Japanese ancestry will diminish the model's ability to highlight the copied segment as exogenous in a recipient genome. Even if LOCO flags a putative introgressed region, any mismatch in the ancestry states can confound the interpretation, raising the question of whether the signal truly represents an introgression or is simply a misassigned region caused by convergence to a suboptimal local maximum during EM optimisation. This uncertainty could result in false positives, where erroneous signals are wrongly considered proof of introgression, and false

negatives, when actual introgressed portions are not detected. Moreover, short introgressions are particularly vulnerable to assignment errors. In Chapter 2, when we simulated a 20-locus introgression, the model failed to detect the introgression. In a slightly longer segment within the 25–50 loci range, the average matching percentage to detect the correct ancestry fo the introgression was only 12.95%, when introgressions spanned 50–75 loci, the matching percentage improved to an average of 60.29%, and for segments between 75 and 100 loci, the average matching reached 93.38%. LOCO's hidden Markov model depends on identifying changes in the recombination and miscopying event pattern. For longer segments, these shifts are easier to identify because there are more data points to support the change. On the other hand, in shorter segments, the recombination signal is frequently too weak given fewer data points, which makes it difficult for the model to distinguish between a real introgression event and a simple data fluctuation. As a result, the model might interpret a short introgressed segment as part of a contiguous block from a single ancestral source. This phenomenon is not unique to LOCO; similar challenges have been observed in other local ancestry models. For instance, Maples *et al.* demonstrated that window-based discriminative methods like RFMix often produce switch errors at recombination boundaries, while Guan reported that short ancestral tracts typically yield lower inference accuracy due to a lack of distinguishing information (Guan, 2014; Maples et al., 2013). Keilwagen *et al.* studied the detection of large-scale introgressions under low-coverage sequencing data (Keilwagen et al., 2023). In their work, they took genebank collections of wheat, aiming to identify known and previously uncharacterised introgressions. They aligned each sample’s reads to a wheat reference and measured coverage for each genomic region. Where a large block of reads mapped poorly compared to the rest of the genome, they flagged a potential introgression. Their findings demonstrated that many landraces and older cultivars carried substantial non-wheat segments acquired through past hybridisation events. Importantly, they showed that ultra-low coverage (ulcWGS) data (i.e., <0.5X) could be sufficient to reveal major introgressions. Their focus was on interspecific introgression in wheat, but these principles are also relevant to detecting second-generation GMO-like modifications in other crops. Keilwagen *et al.* emphasised that detecting introgressions depends heavily

on the size and distinctiveness of the inserted segments. By analogy, local ancestry algorithms such as LOCO could likewise struggle with very short second-generation GMO edits, yet they may succeed in uncovering more extensive genome edits. Moreover, because they demonstrated introgression detection using ultra-low coverage datasets (<0.5×), their work suggests that LOCO could operate effectively with low-quality sequencing data, particularly for larger genome modifications. In Chapter 3, we applied the LOCO model for detecting introgressed segments in second-generation GMO-like genomes. While LOCO shows potential as a screening tool for small genomic modifications in low-coverage data, further methodological improvements are essential to ensure its reliability for practical applications in second-generation GMO detection

Authors contributions

Daniel Wegmann (DW), Daniel Croll (DC) and C. Sarai Reyes-Avila (CSRA) conceived the idea. DW, Madleina Caduff (MC) and C. Sarai Reyes-Avila developed the low-depth copy model applied in this chapter. CSRA implemented the low-depth copy model methods in collaboration with MC. CSRA conducted all data analyses with DC's advice and guidance. CSRA and DC led the writing of this manuscript.

GENERAL DISCUSSION

This thesis examined sequencing and computational methods, in order to identify exogenous DNA specifically in genetically modified organisms (GMOs) and assign ancestry in admixed genomes. In Chapter 1, we developed a microfluidics-based multiplexed amplicon sequencing test for GMO detection. The results included integrating ribosomal and chloroplast markers for species identification, creating and applying 230 amplicons that target GM sequences across diverse crops, and identifying GM sequences even at low concentrations. The study demonstrates how multiplexed sequencing can be used as a high-throughput, scalable substitute for conventional qPCR-based GMO detection methods. This approach can be beneficial in post-market monitoring and regulatory compliance efforts. In Chapter 2, we developed a low-depth copy model (LOCO), which is an ancestry inference algorithm designed to work without predefined reference panels. Building upon Li & Stephens's copy model, LOCO integrates a data-driven haplotype reconstruction strategy similar to that proposed in STITCH. LOCO circumvents the dependency on curated reference panels by directly modeling genotype likelihoods from low-depth sequencing data within a hidden Markov model (HMM) framework. In Chapter 3, we apply an ancestry inference method for screening second-generation GMO-like modifications as a proof-of-principle approach with our low-depth copy model algorithm. Using a rice dataset, we simulated GMO-like modifications through artificial introgressions of 100 SNP and 50 SNP segments.

The standard for detecting GMOs is quantitative PCR, which is labour-intensive. Enhancing and growing detection techniques is essential, given the importance of GMO detection efforts. Our study evaluated the feasibility of a microfluidics-based targeted amplicon sequencing approach by designing a set of 230 amplicons covering 82 unique GM events. Previous studies on multiplex GMO detection have relied on predefined primer sets used in routine food and feed sample screening, based on certified GMO reference materials (Scholtens et al. 2017, Arulandhu et al. 2018). The dependency on a certified set of primers restricts the completeness of the GM event sequences to be recovered because the certified primers do not recover the complete GM event. Our study demonstrates that a *de novo* design of GM amplicons using publicly available sequence information is feasible, and our results show that GM sequences can be

detected in parallel using a microfluidics-based targeted amplicon sequencing strategy. Our assay enables the simultaneous screening of hundreds of samples while allowing for the inclusion of large sets of primers in a single amplification step. The sequence information provides significantly greater certainty about the identity of an amplified sequence in contrast with qPCR approaches that lack direct validation capabilities under non-standard conditions. In the opposite targeted amplicon, sequencing is less sensitive compared to qPCR. In our analyses, the detection of GM sequences at lower concentrations will likely be poorly reproducible. This detection limitation can be partially improved by increasing the overall sequencing coverage of the amplicons, as sensitivity is at least partially correlated with sequencing depth. Our study demonstrates the application of an amplicon sequencing assay in a microfluidics platform that treats samples in parallelisation and the depth of NGS to detect GMOs. Our work fits into efforts to standardise and propose a statistical framework for detecting GMOs (Willems et al. 2016) based on the number of reads aligned per sample. Our workflow expands the capabilities by precisely targeting many sequences of interest and allowing plant species detection in a sample. With 230 designed amplicons corresponding to GM events and additional primers for species-specific barcoding, our approach represents an advancement in GMO detection. We demonstrated the assay robustness by identifying ten known genetic modifications across crop species. It also showed the potential for uncovering undocumented genetic events. The sequence information provided by NGS offers a direct validation of the detected genetic material. This study's successful identification of GMOs underlines the relevance of developing new screening methods. Our analysis shows the relevance of amplicon sequencing, which can be realistically implemented into GMO detection and efficiently analysed using a structured bioinformatics pipeline.

The ancestry inference methods generally rely on pre-defined reference panels. However, predefined reference panels are not always available in cases involving no well-represented populations or ancient DNA (Maples et al., 2013; Price et al., 2009). Our low-depth copy model uses a hidden Markov model framework to infer local ancestry directly from genotype likelihoods without using reference panels. The results indicate that the model can reconstruct diploid ancestry profiles across admixed

genomes. In a simulation test with 10 individuals and 1000 loci, we observed that 98.23% of the diploid ancestry calls inferred correctly matched the simulated values. In another simulation with 20 individuals and 2000 loci, the matching percentage was 98.38%. Across multiple sample sizes and loci combinations, the correct diploid ancestry matching calls ranged from 96.85% to 98.63%. To ensure that our parameters converge to the global maximum, we did our inference tests using parameter values close to the actual simulation values (within 0.0001). In real-world scenarios, the parameters are unknown and must be initialised randomly. The problems arise when the EM becomes stuck in local optima solutions. In the future, we plan to use alternative initialisation methods to help find the global maximum solution (Kwedlo, 2015, You et al., 2023). We can also ask the user to enter the global ancestry proportion parameters for each individual and population as an alternative to help infer the parameters. ADMIXTURE or STRUCTURE can obtain these parameters (Alexander et al., 2009; Pritchard et al., 2000). Inputting the ancestry proportion parameters will help to reduce the parameters to infer, and predicting fewer parameters will lower the chance of suboptimal convergence. The initialisation of the parameters needs to be solved to be able to apply the model to the real-world dataset.

Implementing LOCO to detect second-generation GMO-like modifications by ancestry inference is a new solution for screening genome-edited organisms (GM-LOCO framework). This framework is applicable when genetic material is exchanged between individuals from the same species or closely related individuals. GM-LOCO provides an alternative that does not rely on prior knowledge of the genetic edits. LOCO’s skill to dynamically reconstruct reference panels makes it helpful in detecting subtle genomic modifications without predefined reference datasets, which is needed for identifying undocumented second-generation GMOs that conventional PCR-based approaches that depend on prior knowledge of the sequence can not identify (Chhalliyil et al., 2020). The outcomes of LOCO in the GM-LOCO framework are expected to detect segments in the genomes that show ancestry that deviates from the expected population background, potentially indicating genome modifications. The detection power of LOCO is influenced by the size of the introgression and the population's genetic background. Our current test was done in a dataset of 4000 SNPs. This subset of markers may not capture enough population-specific signals to distinguish between closely related

groups, such as Indonesians and Japanese rice cultivars. Studies in population genetics have demonstrated that the resolution of population structure inference highly depends on the number and the informativeness of the SNPs analysed (Turakulov & Easteal, 2003). Consequently, the limited SNP data in our experiments likely constrained the model's ability to differentiate ancestry sources during parameter estimation. Our study artificially introduced second-generation GMO-like modifications by swapping chromosomal segments between rice populations. The modifications were introduced at a fixed position beyond the 100000 th locus to avoid highly variable regions at the beginning of the chromosome. However, real-world modifications can occur anywhere in the genome, including highly variable regions. Future investigations of LOCO should asses sensitivity across different recombination landscapes. In our experiments for detecting second-generation GMO-like ancestry, inference failed to produce reliable results due to issues with parameter initialisation. Previously, we observed in the LOCO simulations that when the EM algorithm was not initialised with parameters extremely close to the true values, it converged to suboptimal local optima. In our real-world data, however, the true parameters are unknown, making it impossible to initialise them correctly and increasing the risk of suboptimal convergence. This initialisation problem led to inaccurate assignment of ancestry, where individuals were expected to have a predominant Indonesian or Japanese ancestry and were misclassified. Consequently, the detection of introgressed segments was ambiguous. Such misassignments compromise the model's ability to pinpoint genomic regions that deviate from the expected ancestry background, potentially resulting in false positives and false negatives detection in introgression segments. Overcoming the challenges associated with robust parameter initialisation is critical for adapting the GM-LOCO framework for real-world applications (Kwedlo, 2015; You et al., 2023). Assuming we can initialise the parameters correctly using robust strategies, the GM-LOCO framework would still be limited in detecting short genomic segments. Our previous simulations demonstrated that LOCO reliably reconstructs broad diploid ancestry profiles and accurately detects longer introgression segments (e.g., segments spanning 75–100 loci). However, the model's resolution for detecting short segments remains constrained. The underlying reason is that very short segments often do not provide a sufficiently strong signal, given the limited number of SNPs within such regions, to distinguish them from the surrounding genomic background. This limitation has also been observed in other local

ancestry methods (Guan, 2014; Maples et al., 2013). Second-generation GMOs generally have small and targeted modifications, and the relevance of larger introgression detection for second-generation GMOs is uncertain (Grohmann et al., 2019; Kawall, 2019). To contextualise this within the current state of detection methods, most existing approaches focus on detecting specific sequence alterations rather than broader ancestry shifts (Fraiture et al., 2017; Chhalliyil et al., 2020). The ability of GM-LOCO to detect small-scale modifications between closely related individuals remains a key challenge. Our current analysis focused only on Japonica rice variety, Japonica from Indonesia and Japonica from Japan with the introgressed segments swapping between them. Due to the close genetic relationship between these populations, smaller segment swaps may not alter ancestry patterns. GM-LOCO may have the advantage of detecting genetic modifications between closely related individuals. Still, its ability to detect small modifications could decrease when the populations are too genetically similar. Future research should explore detection between still-related populations, such as Japonica and Indica rice varieties, but with more distinct genetic backgrounds. This could help assess the effectiveness of GM-LOCO in identifying smaller modifications in scenarios where genetic differentiation is bigger.

Future work should focus on refining the LOCO-GMO framework to better assess its sensitivity and applicability in detecting small introgressions characteristic of second-generation GMOs. First, LOCO must be able to correctly infer its parameters so that the ancestry inference is trusted. Later, it should be able to be tested under controlled artificial introgression conditions to determine in which conditions and what is the minimal size LOCO can detect in real-world controlled data. We should explore populations that are closely related but still have distinct ancestry patterns given historical separation, such as Japonica and Indica rice varieties. Under these conditions, we can test the performance of LOCO-GMO to detect small introgressions. And better understand the context of the populations where LOCO can or can not detect small introgressions. Once the conditions for detecting small modifications are identified, the next step will be applying LOCO-GMO to real-world second-generation GMOs. Doing that will require working with characterised datasets where individuals' natural recombination patterns are known and gene editing modifications are

documented. Having a real-world, second-generation GMOs data set that is well-characterised can tell if LOCO would be able to distinguish between natural recombination events and artificial modifications introduced by genome editing techniques, validating LOCO-GMO on real-world samples with known genetic modifications will determine its effectiveness as a robust screening tool for identifying genome-edited organisms.

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Arias, A. C. R., P.), A. C.-G. (R I., & López-Pazos, S. A. (2021). A basic scheme of soybean transformation for glyphosate tolerance using Agrobacterium tumefaciens through an approximation of patents: A review. *Agronomía Colombiana*, *39*(2), Article 2. https://doi.org/10.15446/agron.colomb.v39n2.92644

Arulandhu, A. J., van Dijk, J., Staats, M., Hagelaar, R., Voorhuijzen, M., Molenaar, B., van Hoof, R., Li, R., Yang, L., Shi, J., Scholtens, I., & Kok, E. (2018). NGS-based amplicon sequencing approach; towards a new era in GMO screening and detection. *Food Control*, *93*, 201–210. https://doi.org/10.1016/j.foodcont.2018.06.014

Aubry, S., Avila, S. R., Croll, D., & Christ, B. (2021). Chapter 10. Omics-based Detection, Identification and Quantification of GM Food and Feed: Current Challenges and Perspectives. In J. Barros-Velázquez (Ed.), *Food Chemistry, Function and Analysis* (pp. 257–270). Royal Society of Chemistry. https://doi.org/10.1039/9781839163005-00257

Baack, E. J., & Rieseberg, L. H. (2007). A genomic view of introgression and hybrid speciation. *Current Opinion in Genetics & Development*, *17*(6), 513–518. https://doi.org/10.1016/j.gde.2007.09.001

Bilmes, J. A. (n.d.). *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*.

Broeders, S. R. M., De Keersmaecker, S. C. J., & Roosens, N. H. C. (2012). How to Deal with the Upcoming Challenges in GMO Detection in Food and Feed. *Journal of Biomedicine and Biotechnology*, *2012*, 1–11. https://doi.org/10.1155/2012/402418

Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference

Samples. *American Journal of Human Genetics*, *98*(1), 116–126. https://doi.org/10.1016/j.ajhg.2015.11.020

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. A., & Bustamante, C. D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, *107*(2), 786–791. https://doi.org/10.1073/pnas.0909559107

Chen, J., Miao, Z., Kong, D., Zhang, A., Wang, F., Liu, G., Yu, X., Luo, L., & Liu, Y. (2024). Application of CRISPR/Cas9 Technology in Rice Germplasm Innovation and Genetic Improvement. *Genes*, *15*(11), Article 11. https://doi.org/10.3390/genes15111492

Chhalliyil, P., Ilves, H., Kazakov, S. A., Howard, S. J., Johnston, B. H., & Fagan, J. (2020). A Real-Time Quantitative PCR Method Specific for Detection and Quantification of the First Commercialized Genome-Edited Plant. *Foods*, *9*(9), Article 9. https://doi.org/10.3390/foods9091245

Daskalakis, C., Tzamos, C., & Zampetakis, M. (2017). *Ten Steps of EM Suffice for Mixtures of Two Gaussians* (arXiv:1609.00368). arXiv. https://doi.org/10.48550/arXiv.1609.00368

Davies, R. W., Flint, J., Myers, S., & Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, *48*(8), 965–969. https://doi.org/10.1038/ng.3594

Dawkins, R. L., & Lloyd, S. S. (2019). MHC Genomics and Disease: Looking Back to Go Forward. *Cells*, *8*(9), Article 9. https://doi.org/10.3390/cells8090944

Delivery of substances into cells and tissues using a particle bombardment process. (1987). *Particulate Science and Technology*, *5*(1), 27–37. https://doi.org/10.1080/02726358708904533

Divers, J., Palmer, N. D., Langefeld, C. D., Brown, W. M., Lu, L., Hicks, P. J., Smith, S. C., Xu, J., Terry, J. G., Register, T. C., Wagenknecht, L. E., Parks, J. S., Ma, L., Chan, G. C.,

Buxbaum, S. G., Correa, A., Musani, S., Wilson, J. G., Taylor, H. A., … Freedman, B. I. (2017). Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes. *BMC Genetics*, *18*(1), 105. https://doi.org/10.1186/s12863-017-0572-9

Dragoni, F., Garofalo, M., Trotti, R., Liu, Y., Cereda, C., & Gagliardi, S. (2021). Comparison between Conventional qPCR and Microfluidic Chip-Based PCR System for COVID-19 Nucleic Acid Detection. *Journal of Psychiatry and Psychiatric Disorders*, *5*(6), 218–231.

Fraley, C., & Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, *41*(8), 578–588. https://doi.org/10.1093/comjnl/41.8.578

Gaj, T., Gersbach, C. A., & Barbas, C. F. (2013). ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, *31*(7), 397–405. https://doi.org/10.1016/j.tibtech.2013.04.004

Garzón Rodríguez, N., Briceño-Balcázar, I., Nicolini, H., Martínez-Magaña, J. J., Genis-Mendoza, A. D., Flores-Lázaro, J. C., Villatoro Velázquez, J. A., Bustos Gamiño, M., Medina-Mora, M. E., & Quiroz-Padilla, M. F. (2024). Exploring the relationship between admixture and genetic susceptibility to attention deficit hyperactivity disorder in two Latin American cohorts. *Journal of Human Genetics*, *69*(8), 373–380. https://doi.org/10.1038/s10038-024-01246-5

Gene transfer into mouse lyoma cells by electroporation in high electric fields—PubMed. (n.d.). Retrieved February 3, 2025, from https://pubmed.ncbi.nlm.nih.gov/6329708/

GMOs—European Commission. (n.d.). Retrieved March 17, 2025, from https://joint-research-centre.ec.europa.eu/scientific-activities-z/gmos_en

Gomez, F., Wang, L., Abel, H., Zhang, Q., Province, M. A., & Borecki, I. B. (2015). Admixture mapping of coronary artery calcification in African Americans from the NHLBI family heart study. *BMC Genetics*, *16*(1), 42. https://doi.org/10.1186/s12863-015-0196-x

Grohmann, L., Keilwagen, J., Duensing, N., Dagand, E., Hartung, F., Wilhelm, R., Bendiek, J., & Sprink, T. (2019). Detection and Identification of Genome Editing in Plants: Challenges and Opportunities. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00236

Guan, Y. (2014). Detecting Structure of Haplotypes and Local Ancestry. *Genetics*, *196*(3), 625–642. https://doi.org/10.1534/genetics.113.160697

Gupta, R. M., & Musunuru, K. (2014). Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. *The Journal of Clinical Investigation*, *124*(10), 4154–4161. https://doi.org/10.1172/JCI72992

Hassan, S., Surakka, I., Taskinen, M.-R., Salomaa, V., Palotie, A., Wessman, M., Tukiainen, T., Pirinen, M., Palta, P., & Ripatti, S. (2021). High-resolution population-specific recombination rates and their effect on phasing and genotype imputation. *European Journal of Human Genetics*, *29*(4), 615–624. https://doi.org/10.1038/s41431-020-00768-8

Holst-Jensen, A., Spilsberg, B., Arulandhu, A. J., Kok, E., Shi, J., & Zel, J. (2016). Application of whole genome shotgun sequencing for detection and characterization of genetically modified organisms and derived products. *Analytical and Bioanalytical Chemistry*, *408*(17), 4595–4614. https://doi.org/10.1007/s00216-016-9549-1

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., & Zhang, F. (n.d.). *DNA targeting specificity of RNA-guided Cas9 nucleases*. Retrieved March 17, 2025, from https://stacks.cdc.gov/view/cdc/29899

Keilwagen, J., Lehnert, H., Badaeva, E. D., Özkan, H., Sharma, S., Civáň, P., & Kilian, B.

(2023). Finding needles in a haystack: Identification of inter-specific introgressions in wheat genebank collections using low-coverage sequencing data. *Frontiers in Plant Science*, *14*. https://doi.org/10.3389/fpls.2023.1166854

Kleckner, N. (2006). Chiasma formation: Chromatin/axis interplay and the role(s) of the synaptonemal complex. *Chromosoma*, *115*(3), 175–194. https://doi.org/10.1007/s00412-006-0055-7

Koller, D., Wendt, F. R., Pathak, G. A., De Lillo, A., De Angelis, F., Cabrera-Mendoza, B., Tucci, S., & Polimanti, R. (2022). Denisovan and Neanderthal archaic introgression differentially impacted the genetics of complex traits in modern populations. *BMC Biology*, *20*(1), 249. https://doi.org/10.1186/s12915-022-01449-2

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Korunes, K. L., & Goldberg, A. (2021). Human genetic admixture. *PLOS Genetics*, *17*(3), e1009374. https://doi.org/10.1371/journal.pgen.1009374

Kwedlo, W. (2015). A new random approach for initialization of the multiple restart EM algorithm for Gaussian model-based clustering. *Pattern Anal. Appl.*, *18*(4), 757–770. https://doi.org/10.1007/s10044-014-0441-3

Lehnert, T., & M. Gijs, M. A. (2024). Microfluidic systems for infectious disease diagnostics. *Lab on a Chip*, *24*(5), 1441–1493. https://doi.org/10.1039/D4LC00117F

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. https://doi.org/10.48550/arXiv.1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination

hotspots using single-nucleotide polymorphism data. *Genetics*, *165*(4), 2213–2233. https://doi.org/10.1093/genetics/165.4.2213

Lu, K., Wu, B., Wang, J., Zhu, W., Nie, H., Qian, J., Huang, W., & Fang, Z. (2018). Blocking amino acid transporter OsAAP3 improves grain yield by promoting outgrowth buds and increasing tiller number in rice. *Plant Biotechnology Journal*, *16*(10), 1710–1722. https://doi.org/10.1111/pbi.12907

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Marmiroli, N., Maestri, E., Gullì, M., Malcevschi, A., Peano, C., Bordoni, R., & De Bellis, G. (2008). Methods for detection of GMOs in food and feed. *Analytical and Bioanalytical Chemistry*, *392*(3), 369–384. https://doi.org/10.1007/s00216-008-2303-6

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Mizuno, F., Naka, I., Ueda, S., Ohashi, J., & Kurosaki, K. (2021). The number of SNPs required for distinguishing Japanese from other East Asians. *Legal Medicine*, *49*, 101849. https://doi.org/10.1016/j.legalmed.2021.101849

Mosca, M. J., & Cho, H. (2023). Reconstruction of private genomes through reference-based genotype imputation. *Genome Biology*, *24*(1), 271. https://doi.org/10.1186/s13059-023-03105-6

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, N.Y.)*, *310*(5746), 321–324. https://doi.org/10.1126/science.1117196

Neff, W. E., Mounts, T. L., Rinsch, W. M., Konishi, H., & El-Agaimy, M. A. (1994). Oxidative

stability of purified canola oil triacylglycerols with altered fatty acid compositions as affected by triacylglycerol composition and structure. *Journal of the American Oil Chemists' Society*, *71*(10), 1101–1109. https://doi.org/10.1007/BF02675903

Ozyigit, I. I. (2020). Gene transfer to plants by electroporation: Methods and applications. *Molecular Biology Reports*, *47*(4), 3195–3210. https://doi.org/10.1007/s11033-020-05343-4

Padhukasahasram, B. (2014). Inferring ancestry from population genomic data and its applications. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00204

Pankratov, V., Mezzavilla, M., Aneli, S., Kuznetsov, I. A., Fusco, D., Wilson, J. F., Metspalu, M., Provero, P., Pagani, L., & Marnetto, D. (2024). Ancestral genetic components are consistently associated with the complex trait landscape in European biobanks. *European Journal of Human Genetics*, 1–8. https://doi.org/10.1038/s41431-024-01678-9

Paul, J.-Y., Khanna, H., Kleidon, J., Hoang, P., Geijskes, J., Daniells, J., Zaplin, E., Rosenberg, Y., James, A., Mlalazi, B., Deo, P., Arinaitwe, G., Namanya, P., Becker, D., Tindamanyire, J., Tushemereirwe, W., Harding, R., & Dale, J. (2017). Golden bananas in the field: Elevated fruit pro-vitamin A from the expression of a single banana transgene. *Plant Biotechnology Journal*, *15*(4), 520–532. https://doi.org/10.1111/pbi.12650

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., & Myers, S. (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLOS Genetics*, *5*(6), e1000519. https://doi.org/10.1371/journal.pgen.1000519

Priestley, A. L., & Brownbridge, M. (2009). Field trials to evaluate effects of Bt-transgenic silage corn expressing the Cry1Ab insecticidal toxin on non-target soil arthropods in northern New England, USA. *Transgenic Research*, *18*(3), 425–443. https://doi.org/10.1007/s11248-008-9234-z

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using

Multilocus Genotype Data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Program, H. F. (2024). Science and History of GMOs and Other Food Modification Processes. *FDA*. https://www.fda.gov/food/agricultural-biotechnology/science-and-history-gmos-and-other-food-modification-processes

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Salter-Townshend, M., & Myers, S. (2019). Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics*, *212*(3), 869–889. https://doi.org/10.1534/genetics.119.302139

Saltykova, A., Van Braekel, J., Papazova, N., Fraiture, M.-A., Deforce, D., Vanneste, K., De Keersmaecker, S. C. J., & Roosens, N. H. (2022). Detection and identification of authorized and unauthorized GMOs using high-throughput sequencing with the support of a sequence-based GMO database. *Food Chemistry: Molecular Sciences*, *4*, 100096. https://doi.org/10.1016/j.fochms.2022.100096

Sander, J. D., & Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, *32*(4), 347–355. https://doi.org/10.1038/nbt.2842

Sankararaman, S., Sridhar, S., Kimmel, G., & Halperin, E. (2008). Estimating Local Ancestry in Admixed Populations. *American Journal of Human Genetics*, *82*(2), 290–303. https://doi.org/10.1016/j.ajhg.2007.09.022

Scholtens, I. M. J., Molenaar, B., van Hoof, R. A., Zaaijer, S., Prins, T. W., & Kok, E. J. (2017). Semiautomated TaqMan PCR screening of GMO labelled samples for (unauthorised) GMOs. *Analytical and Bioanalytical Chemistry*, *409*(15), 3877–3889. https://doi.org/10.1007/s00216-017-0333-7

Schwenk, K., Brede, N., & Streit, B. (2008). Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1505), 2805–2811. https://doi.org/10.1098/rstb.2008.0055

Shriner, D. (2013). Overview of Admixture Mapping. *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]*, *CHAPTER*, Unit1.23. https://doi.org/10.1002/0471142905.hg0123s76

The 3, 000 rice genomes project. (2014). The 3,000 rice genomes project. *GigaScience*, *3*(1), 7. https://doi.org/10.1186/2047-217X-3-7

The Production of Mutations by X-Rays—PMC. (n.d.). Retrieved February 3, 2025, from https://pmc.ncbi.nlm.nih.gov/articles/PMC1085688/

Tiwari, J. K., Singh, A. K., & Behera, T. K. (2023). CRISPR/Cas genome editing in tomato improvement: Advances and applications. *Frontiers in Plant Science*, *14*. https://doi.org/10.3389/fpls.2023.1121209

Tripathi, J. N., Ntui, V. O., Ron, M., Muiruri, S. K., Britt, A., & Tripathi, L. (2019). CRISPR/Cas9 editing of endogenous banana streak virus in the B genome of Musa spp. Overcomes a major challenge in banana breeding. *Communications Biology*, *2*(1), 46. https://doi.org/10.1038/s42003-019-0288-7

Turakulov, R., & Easteal, S. (2003). Number of SNPS Loci Needed to Detect Population Structure. *Human Heredity*, *55*(1), 37–45. https://doi.org/10.1159/000071808

Wang, B., Li, N., Huang, S., Hu, J., Wang, Q., Tang, Y., Yang, T., Asmutola, P., Wang, J., & Yu, Q. (2021). Enhanced soluble sugar content in tomato fruit using CRISPR/Cas9-mediated SlINVINH1 and SlVPE5 gene editing. *PeerJ*, *9*, e12478. https://doi.org/10.7717/peerj.12478

Wang, F., Fan, F., Li, W., Zhu, J., Wang, J., Zhong, W., & Yang, J. (2016). Knock-out Efficiency Analysis of Pi21 Gene Using CRISPR/Cas9 in Rice. *Chinese Journal OF Rice Science*,

30(5), 469. https://doi.org/10.16819/j.1001-7216.2016.6009

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C., & Qiu, J.-L. (2014). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nature Biotechnology*, *32*(9), 947–951. https://doi.org/10.1038/nbt.2969

Wegmann, D., Kessner, D. E., Veeramah, K. R., Mathias, R. A., Nicolae, D. L., Yanek, L. R., Sun, Y. V., Torgerson, D. G., Rafaels, N., Mosley, T., Becker, L. C., Ruczinski, I., Beaty, T. H., Kardia, S. L. R., Meyers, D. A., Barnes, K. C., Becker, D. M., Freimer, N. B., & Novembre, J. (2011). Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics*, *43*(9), 847–853. https://doi.org/10.1038/ng.894

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001-1006. https://doi.org/10.1093/nar/gkt1229

Wu, Y., & Zhou, H. H. (2019). *Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations* (arXiv:1908.10935). arXiv. https://doi.org/10.48550/arXiv.1908.10935

Xia, R., Jian, X., Rodrigue, A. L., Bressler, J., Boerwinkle, E., Cui, B., Daviglus, M. L., DeCarli, C., Gallo, L. C., Glahn, D. C., Knowles, E. E. M., Moon, J.-Y., Mosley, T. H., Satizabal, C. L., Sofer, T., Tarraf, W., Testai, F., Blangero, J., Seshadri, S., … Fornage, M. (2024). Admixture mapping of cognitive function in diverse Hispanic and Latino adults: Results from the Hispanic Community Health Study/Study of Latinos. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*. https://doi.org/10.1002/alz.14082

Ye, X., Al-Babili, S., Klöti, A., Zhang, J., Lucca, P., Beyer, P., & Potrykus, I. (2000). Engineering the Provitamin A (β-Carotene) Biosynthetic Pathway into (Carotenoid-Free) Rice Endosperm. *Science*, *287*(5451), 303–305. https://doi.org/10.1126/science.287.5451.303

You, J., Li, Z., & Du, J. (2023). A new iterative initialization of EM algorithm for Gaussian mixture models. *PLOS ONE*, *18*(4), e0284114. https://doi.org/10.1371/journal.pone.0284114

Yuan, K., Zhou, Y., Ni, X., Wang, Y., Liu, C., & Xu, S. (2017). Models, methods and tools for ancestry inference and admixture analysis. *Quantitative Biology*, *5*(3), 236–250. https://doi.org/10.1007/s40484-017-0117-2

Zambryski, P., Joos, H., Genetello, C., Leemans, J., Montagu, M. V., & Schell, J. (1983). Ti plasmid vector for the introduction of DNA into plant cells without alteration of their normal regeneration capacity. *The EMBO Journal*, *2*(12), 2143–2150. https://doi.org/10.1002/j.1460-2075.1983.tb01715.x

Zeng, X., Luo, Y., Vu, N. T. Q., Shen, S., Xia, K., & Zhang, M. (2020). CRISPR/Cas9-mediated mutation of OsSWEET14 in rice cv. Zhonghua11 confers resistance to Xanthomonas oryzae pv. Oryzae without yield penalty. *BMC Plant Biology*, *20*(1), 313. https://doi.org/10.1186/s12870-020-02524-y

Zhang, A., Liu, Y., Wang, F., Li, T., Chen, Z., Kong, D., Bi, J., Zhang, F., Luo, X., Wang, J., Tang, J., Yu, X., Liu, G., & Luo, L. (2019). Enhanced rice salinity tolerance via CRISPR/Cas9-targeted mutagenesis of the OsRR22 gene. *Molecular Breeding*, *39*(3), 47. https://doi.org/10.1007/s11032-019-0954-y

Zhao, D.-S., Li, Q.-F., Zhang, C.-Q., Zhang, C., Yang, Q.-Q., Pan, L.-X., Ren, X.-Y., Lu, J., Gu, M.-H., & Liu, Q.-Q. (2018). GS9 acts as a transcriptional activator to regulate rice grain shape and appearance quality. *Nature Communications*, *9*, 1240. https://doi.org/10.1038/s41467-018-03616-y

PUBLICATIONS

Detection of genetically modified organisms using highly multiplexed amplicon sequencing. Reyes-Avila, C. S., Waldvogel, D., Pradervand, N., Aubry, S., & Croll, D. Food Control, 2024.

Genome Evolution in Fungal Plant Pathogens: From Populations to Kingdom-Wide Dynamics. Ursula Oggenfuss, Alice Feurtey, Claudia Sarai Reyes-Avila, Emile Gluck-Thaler, Guido Puccetti, Hanna Maren Glad, Leen Nanchira Abraham, Luzia Stalder, Sabina Moser Tralamazza, Sandra Milena González-Sáyer & Daniel Croll. Part of the The Mycota book series MYCOTA,volume 14, 2023.

Histone H3K27 Methylation Perturbs Transcriptional Robustness and Underpins Dispensability of Highly Conserved Genes in Fungi. Sabina Moser Tralamazza, Leen Nanchira Abraham, Claudia Sarai Reyes-Avila, Benedito Corrêa, Daniel Croll. Molecular Biology And Evolution, 2021.

Chapter 10. Omics-based Detection, Identification and Quantification of GM Food and Feed: Current Challenges and Perspectives. Sylvain Aubry, Sarai Reyes Avila, Daniel Croll, Bastien Christ. Food Chemistry, Function and Analysis, 2021.

CV

Sarai Reyes-Avila

ORCID ID: https://orcid.org/0000-0002-6970-8862

EDUCATION

Bachelor's degree in Genomic Sciences
Institute of Biotechnology, National Autonomous University of Mexico (Spanish: UNAM)
Cuernavaca, Morelos, México
GPA: 9.29 out of 10
Thesis project: Metagenomic analysis of common vampire bats and other bioinformatics applications.
Advisor: Professor Tom Gilbert at Natural History Museum of Denmark
Dates: August 2015 to July 2019

RESEARCH EXPERIENCE

PhD student at the Faculty of Science, University of Neuchatel
Section: Evolutionary Genomics of Pathogens
Supervisor: Professor Daniel Croll
Location: Neuchatel, Switzerland
Period: November 2020 to present
Thesis Project: **Detection And Characterization Of Exogenous DNA. From Genetically Modified Organisms (GMOs) to Naturally Admixed Genomes.**

Research Assistant at the Faculty of Health and Medical Science, University of Copenhagen
Institute: The GLOBE Institute, before: Natural History Museum of Denmark
Section: Evolutionary Genomics
Supervisor: Professor Tom Gilbert
Location: Copenhagen, Denmark
Period: October 2018 to July 2020
Project 1: **Metagenomics** analysis of blood meal from common vampire bats to identify prey and microbios diversity.
Project 2. Develop a bioinformatics pipeline to process reduced representation libraries of degraded DNA, from **ancient human DNA** and wolves feces DNA.
Project 3: Study the behavior of **extinct species** based on the hypothesis hominid archaic humans extinct because their reproduction was lack of compatibility.

Visiting student at Max-Planck-Campus Tübingen
Department: Molecular Biology
Supervisor: Dr. Hernan Burbano
Location: Tubingen, Germany
Period: May 2018 to May 2018 (3 weeks)
Project: Visit our collaborator to perform population genetics analysis of *Magnaporthe oryzae* or blast disease of crops.

Pre-doc Student at The Sainsbury Laboratory, University of East Anglia
Subject: Evolutionary molecular plant-microbe interactions
Supervisor: Professor Sophien Kamoun and Dr. Joe Win
Location: Norwich, United Kingdom
Period: September 2017 to August 2018
Project: Evolutionary analysis of *Magnaporthe oryzae* or blast disease in crops.

Summer Internship at Helmholtz Zentrum München
Department: Institute of Computational Biology
Supervisor: Professor Matthias Heinig
Location: Munich, Germany
Period: June 2017 to August 2017
Project: Integrated computational analysis of DNA methylation and gene expression.

Undergraduate Research Assistant at Center for Genomics Sciences of National Autonomous University of Mexico (UNAM in Spanish)
Department: Genome Ecology
Supervisor: Professor Esperanza Martinez
Location: Cuernavaca, Mexico
Period: January 2015 to July 2015
Project: Study of the symbiosis between *Oscheius tipulae* (Nematode) -*Rhizobium etli* (bacteria) -*Phaseolus vulgaris* (plant). My project sought to understand what is the relationship of the nematode with the bacteria and how this relationship affected the growth of the plant.

PUBLICATIONS

Reyes-Avila, C. S., Waldvogel, D., Pradervand, N., Aubry, S., & Croll, D. Detection of genetically modified organisms using highly multiplexed amplicon sequencing. *Food Control* (2024). https://www.sciencedirect.com/science/article/pii/S0956713524003876

Ursula Oggenfuss, Alice Feurtey, **Claudia Sarai Reyes-Avila**, Emile Gluck-Thaler, Guido Puccetti, Hanna Maren Glad, Leen Nanchira Abraham, Luzia Stalder, Sabina Moser Tralamazza, Sandra Milena González-Sáyer, Daniel Croll. Genome Evolution in Fungal Plant Pathogens: From Populations to Kingdom-Wide Dynamics. *The Mycota*, volume 14 (2023). https://link.springer.com/chapter/10.1007/978-3-031-29199-9_5

Sergio M. Latorre, Vincent M. Were, Andrew J. Foster, Thorsten Langner, Angus Malmgren, Adeline Harant, Soichiro Asuke, **Sarai Reyes-Avila**, … Joe Win, Nicholas J. Talbot, Hernán A. Burbano, Sophien Kamoun. Convergent loss of an effector gene during adaptation of *Magnaporthe oryzae* to finger millet. *PLOS Biology* (2023). https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3002052

Sabina Moser Tralamazza, Leen Nanchira Abraham, **Claudia Sarai Reyes-Avila**, Benedito Corrêa, Daniel Croll. Histone H3K27 Methylation Perturbs Transcriptional Robustness and Underpins Dispensability of Highly Conserved Genes in Fungi. *Molecular Biology And Evolution* (2021). https://academic.oup.com/mbe/article/39/1/msab323/6424003?login=true

Sylvain Aubry, **Sarai Reyes Avila**, Daniel Croll, Bastien Christ. Chapter 10. Omics-based Detection, Identification and Quantification of GM Food and Feed: Current Challenges and Perspectives. *Food Chemistry, Function and Analysis* (2021). https://books.rsc.org/books/edited-volume/887/chapter-abstract/669810/Omics-based-Detection-Identification-and

Physilia Ying Shi Chua, Christian Carøe, Alex Crampton-Platt, **Claudia Sarai Reyes-Avila**, Gareth Jones, Daniel G. Streicker, Kristine Bohmann. A two-step metagenomics approach for prey identification from the blood meals of common vampire bats (*Desmodus rotundus*). Metabarcoding and Metagenomics (2022) https://mbmg.pensoft.net/article/78756/

Sergio Latorre, **Reyes-Avila C Sarai**, Joe Win, Sophien Kamoun S, Hernan Burbano. Differential loss of effector genes in three recently expanded pandemic clonal lineages of the rice blast fungus. BMCBiology (2020). https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-020-00818-z

Joe Win, Emilie Chanclud, **Reyes-Avila C Sarai**, Thorsten Langner, Islam MT, Sophien Kamoun. Nanopore sequencing of genomic DNA from *Magnaporthe oryzae* isolates from different hosts. Zenodo (2019). http://doi.org/10.5281/zenodo.2564950.

C. Sarai Reyes-Avila, Elisa Nunez. Edición Genética en Medicina. NOTA-INCyTU (2018). http://www.foroconsultivo.org.mx/INCyTU/documentos/Completa/INCyTU_18-010.pdf

Gupta DR, **Reyes-Avila CS**, Win J, Soanes DM, Ryder LS, Croll D, Bhattacharjee P, Hossain MS, Mahmud NU, Mehebub MS, Surovy MZ, Rahman MM, Talbot N, Kamoun S, Islam MT. Cautionary Notes on Use of the MoT3 Diagnostic Assay for Magnaporthe oryzae Wheat and Rice Blast Isolates. Phytopathology (2017). https://www.ncbi.nlm.nih.gov/pubmed/30253117

Acknowledgements

First, I want to thank my supervisor, Daniel. Thanks for gave me the opportunity to join your laboratory. His support allowed me to explore and develop projects that I was interested about. Thanks to Sylvain, Nicolas and Dominique for their help in the development of our GMO detection assay. Especially thank Sylvain for his mentoring beyond the project. Thanks to Daniel W. and Madelina for their help in the mathematical formalisation and implementation of our low-depth copy model. Particularly to Madliena for her inmense help in every detail of the project. A deep thank to the entire evolutionary genomics team for creating an amazing work environment. A special mention to Sabina and Ursula for their scientific insight and support. Lastly, I would like to express deep gratitude to my family. You are my motivation. Sorry for my absence but thank for your love and support.