BMC Genomics

RESEARCH Open Access

HASCH - A high-throughput amplicon-based SNP-platform for medicinal cannabis and industrial hemp genotyping applications

Locedie Mansueto[1], Erwin Tandayu[1], Jos Mieog[1], Lennard Garcia-de Heer[1], Rekhamani Das[1], Adam Burn[1], Ramil Mauleon[1,2] and Tobias Kretzschmar[1*]

Abstract

Background *Cannabis sativa* is seeing a global resurgence as a food, fiber and medicinal crop for industrial hemp and medicinal Cannabis industries respectively. However, a widespread moratorium on the use and research of *C. sativa* throughout most of the 20th century has seen the development of improved cultivars for specific end uses lag behind that of conventional crops. While *C. sativa* research and development has seen significant investments in the recent past, resulting in a suite of publicly available genomic resources and tools, a versatile and cost-effective mid-density genotyping platform for applied purposes in breeding and pre-breeding is lacking. Here we report on a first mid-density fixed-target SNP platform for *C. sativa*.

Results The High-throughput Amplicon-based SNP-platform for medicinal Cannabis and industrial Hemp (HASCH) was designed using a combination of filtering and Integer Linear Programming on publicly available whole-genome sequencing and RNA sequencing data, supplemented with in-house generated genotyping-by-sequencing (GBS) data. HASCH contains 1,504 genome-wide targets of high call rate (97% mean) and even distribution across the genome, designed to be highly informative (> 0.3 minor allele frequency) across both medicinal cannabis and industrial hemp gene pools. Average numbers of mismatch SNP between any two accessions were 251 for medicinal cannabis ($N = 116$) and 272 for industrial hemp ($N = 87$). Comparing HASCH data with corresponding GBS data on a collection of diverse *C. sativa* accessions demonstrated high concordance and resulted in comparable phylogenies and genetic distance matrices. Using HASCH on a segregating F2 population derived from a cross between a tetrahydrocannabinol (THC)-dominant and a cannabidiol (CBD)-dominant accession resulted in a genetic map consisting of 310 markers, comprising 10 linkage groups and a total size of 582.7 cM. Quantitative Trait Locus (QTL) mapping identified a major QTL for CBD content on chromosome 7, consistent with previous findings.

Conclusion HASCH constitutes a versatile, easy to use and cost-effective genotyping solution for the rapidly growing Cannabis research community. It provides consistent genetic fingerprints of 1504 SNPs with wide applicability genetic resource management, quantitative genetics and breeding.

*Correspondence:
Tobias Kretzschmar
Tobias.Kretzschmar@scu.edu.au

Full list of author information is available at the end of the article

Keywords Cannabis sativa, Genotyping, SNP, Plant breeding, Quantitative trait loci, Integer linear programming, Genetic fingerprints

Introduction

Cannabis sativa is an underutilized crop of high potential for food, fiber and medicinal markets. Female *C. sativa* flowers produce an abundance of terpenoids and cannabinoids, secondary compounds that are increasingly appreciated for nutraceutical and medicinal applications [1–3]. *C. sativa* seeds are rich in polyunsaturated fatty acids and essential amino acids, making it a promising oil and protein crop [3, 4]. In addition, *C. sativa* is a prolific biomass accumulator that produces high-value bast and hurd fibers, suitable for textiles and carbon-negative construction materials.

However, despite its versatility and being considered one of the earliest crops, *C. sativa* is still regarded as only semi-domesticated, lagging in development in respect to the extent of genetic improvement achieved for conventional crops [5]. Modern breeding approaches to improve *C. sativa* for key domestication traits, specific end-uses and varying environments and cultivation practices are still in their infancy. This is largely due to tight regulation and prohibition, which limited research and development of *C. sativa* for the better part of the last century. While these restrictions have been easing recently, *C. sativa* has missed out on the green revolution and subsequent molecular breeding technologies that have greatly enhanced the performance and adaptation of conventional crops.

Advances in genotyping, and the application of single nucleotide polymorphisms (SNPs) in particular, have significantly accelerated crop improvement over the last decades [6]. SNPs are abundant, biallelic, co-dominant and evenly distributed along the genome, enabling rapid, routine and cost-effective genotyping solutions ranging from targeted single trait marker selection [7, 8] to high density genome-wide marker applications [9–12]. In pre-breeding, SNP platforms facilitate genetic diversity studies, Quantitative Trait Locus (QTL) discovery and QTL introgression. In breeding, SNP platforms enable integrating marker assisted selection (MAS), backcrossing programs and genomic selection (GS) approaches. Moreover, SNPs contribute to increased efficiencies of quality control measures such as varietal identification, pedigree verification and seed purity assessment [13].

Next Generation Sequencing (NGS) and array-based technologies are the two dominant SNP detection systems, though amplicon based-systems have recently gained in importance [14, 15].

As a result of the recent resurgence in hemp and medicinal cannabis research and development, *C. sativa* reference genomes are now publicly available for Finola, a hemp seed variety, Purple Kush, a medicinal cultivar, and CBDRx [16–18], as well as a wealth of other high-density genotyping information. Untargeted NGS platforms, such as Genotyping-By-Sequencing (GBS) and RNA sequencing, have been used in *C. sativa* genetic diversity studies and have started to be utilized in quantitative genetic studies as well. Lynch et al. [19] analyzed 340 diverse varieties and demonstrated the existence of at least three major groups. McKernan et al. [20] sequenced 42 genomes and revealed extensive copy number variation in cannabinoid biosynthesis and pathogen resistance genes. Ren et al. [21] sequenced 110 accessions from worldwide origins and showed that *C. sativa* was first domesticated in East Asia, and current cultivars diverged from an ancestral gene pool represented by feral plants and landraces in China. Woods et al. [22] sequenced diverse samples of feral and domesticated lineages of *C. sativa* from U.S. and German (Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)) collections. In addition to published journal articles, commercial groups like Kannapedia [23] and Phylos (NCBI projects PRJNA347566, PRJNA510566), also provide publicly available sequences. RNA-Seq data from trichome of medicinal samples are available from studies investigating associations between transcript expression and metabolite abundance [24, 25]. In addition, Livingston et al. [26] generated RNA-Seq for Finola trichomes and other transcriptome data on for a wide range of tissues is publicly available [27].

While resequencing, GBS and RNA sequencing for SNP calling do not require upfront platform development and minimize ascertainment bias typically associated with targeted platforms, they require complex experimental protocols, sophisticated data analysis, and bioinformatics pipelines to process raw sequence data into useful genotypic matrices, adding to the cost and time of genotyping. This added cost in combination with a relatively high cost per sample currently limits the applicability of NGS platforms in breeding programs and pre-breeding applications such as QTL mapping [28].

Despite of a wealth of high-density genotyping data for *C. sativa* being in the public domain, little effort has been made in utilizing the underlying variant information for the development of targeted genotyping solutions for more applied purposes such as (pre-) breeding. Only a small number of QTL mapping studies have been published to date, covering a range of traits in hemp and medicinal Cannabis and industrial hemp [18, 29–31]. They all relied on high-density genotyping to generate genetics maps and establish marker trait associations. While high-resolution maps of high marker density are of advantage for certain applications, they tend to be overkill for the mapping of QTL in segregating populations, where frequency of recombination rather than marker density is the limiting factor [32].

Large-scale applications in *C. sativa* breeding with emphasis on population improvement necessitate routine genotyping of thousands of lines per season at the shortest turnaround time possible to make in-season decisions. The current complexity of GBS technologies are limiting in this context.

Here, we report the development of a **H**igh-throughput **A**mplicon-based **S**NP-platform for medicinal **C**annabis and industrial **H**emp – HASCH -, for the cost-effective amplification and sequencing of 1,504 highly informative SNP sites in a 384-plex protocol that has high accuracy and high efficiency. We demonstrate the applicability of HASCH for diversity studies, the establishment of genetic maps and QTL mapping approaches.

Materials and methods

Plant materials

A total of 376 samples were used in his study (Supplementary file 1) derived from the following plant materials.

- **IPK Genebank collection**: 156 individuals from 55 accessions, with 2–4 individuals per accession. These non-proprietary genebank accessions are publicly available from the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstraße 3, 06466 Gatersleben, Germany.
- **Australian private collection**: 36 individuals from 26 accessions, with four individuals from No. 8 and 8 individuals from Futura. These accessions are proprietary materials from Kavasil Pty LTd (ABN 24 150 989 629), Nimbin, NSW 2480, Australia.
- **Chinese commercial varieties**: 4 single individuals from 4 accessions. These accessions are commercially available from The Hemp Corporation Pty Ltd, Hay, NSW, 2711, Australia.
- **3 repeat control samples**: derived from the IPK Genebank collection.
- **F2 mapping population**: 153 lines of a feminized F2 population. A female CBD-dominant individual of IPK_57 was crossed with a male THC-dominant individual of IPK_36. A single female F1 progeny (of intermediate CBD/THC content) was selected and self-fertilized using STS treatment (see below) to generate a segregating F2 population.
- **Inbred lines**: 24 samples from 6 inbred lines – F1- F4, derived from either the "IPK Genebank collection" or the "Australian private collection".

C. sativa cultivation, sample, storage, processing of plant material and cannabinoid analysis were performed in strict adherence to Sect. 23(4)(b) and 41(b) of the NSW Drug Misuse and Trafficking Act 1985, held under the Authority granted to Prof. Bronwyn Barkla of Southern Cross University, issued by the New South Wales Ministry of Health, Australia. The private "Australian private collection" was received and handled under a bilateral Material Transfer Agreement (MTA) between Southern Cross University and Kavasil Pty Ltd, while the "IPK genebank collection" was received and handled under the FAO governed Standard Material Transfer Agrement (sMTA) (https://www.fao.org/plant-treaty/areas-of-work/the-multilateral-system/the-smta/en/). The "IPK genebank collection" was imported from Germany to Australia under a federal Office of Drug Control (ODC) license to import No. 1,820,928.

Plant cultivation and sampling

Seeds were germinated on folded wet paper towels soaked in a 0.3% hydrogen peroxide solution and sealed in plastic ziplock bags for 1–2 days, then transferred onto fresh paper towel soaked in water and returned to the ziplock for another 2–4 days. Germinating seeds were transferred into seedling trays with wet 70% cocopeat, 30% perlite potting mix with pH 6. Alternatively, seeds were directly germinated in seedling trays with the potting mix. Half-strength Canna coco (CANNA) was sprayed on and around the seedlings in the trays every other day. Once the second pair of true leaves became visible, the seedlings were transferred to pots for further growth and development.

All accessions, but not the 153 lines of the F2 population, were grown in pollen proof growth chambers. Seedlings were transferred to 9 L pots (4–5 seedlings/pot) with the same potting mix as before but supplemented with 45 g of Osmocote exact standard 3–4 months (ICL) and 45 g of Osmocote exact standard 5–6 months. After leaf samples were taken for DNA, each accession was sealed in a separate pollen-proof chamber, which prevented any cross-fertilisations between accessions. Plants were grown under long-day conditions (18 H L, 6 H D) until about 25 cm high, when flowering was induced by switching to short-day conditions (12 h L, 12 H D). Watering was automatic via periodic flooding of the chamber base (1–2 cm). Plants were unsealed and harvested once at least 70% of seed was considered matured. An inflorescence sample was obtained during seed harvest, which was dried before being ground into a powder for GC-MS analysis.

For inbreeding, the sex of the seedlings was determined using the MADC2 primers under previously published conditions and female individuals were then selected for self-fertilisation [33]. Male flowers were then induced on a single branch through the application of 6mM silver thiosulfate (STS) every second day for five days once the photoperiod was reduced to 12:12, for a total of three ~ 10 ml treatments [34].

The F2 population was generated from a controlled cross between tetrahydrocannabinol (THC)-dominant IPK_CAN_36 (Male) and cannabidiol (CBD)-dominant IPK_CAN_57 (Female). A selected female F1 individual was self-fertilized using STS treatment and the resulting all-female F2 population was grown as described above with minor modification. Seeds were directly sown into 0.5 L

pots on trays with constant watering system under control conditions. Temperatures ranged from 27±2° C and plants were grown for three weeks under long-day conditions (18 H L, 6 H D) before switching to short-day conditions (12 h L, 12 H D). Young, fully expanded leaves were collected and stored at -20 ° C for DNA extraction. Apical flowers were collected and dried at 15° C and 15% relative humidity for subsequent CBD and THC quantification.

DNA extraction

Genomic DNA (gDNA) was extracted from leaf tissue of single plants using a Qiagen Plant mini kit. DNA integrity and quality was checked visually on 1% agarose gel, while DNA quantity was assessed using Qubit 2.0 (*Thermo Fisher Scientific*) fluorometric kits. DNA digestibility was checked using HindIII.

Genotyping by sequencing and SNP calling

The genotyping-by-sequencing data was generated following established methodology [35], with the following changes: 100ng of genomic DNA were used, 3.6 ng of total adapters were used, the genomic DNAs were restricted with ApeKI enzyme and the library was amplified with 18 PCR cycles. Sequencing results were processed by demultiplexing using axe-demux [36] then trimmed using GBS-PreProcess [37]. SNP calling was performed using TASSEL-GBS pipeline [38] against the CBDRx assembly. The samples used for GBS are listed in Supplementary file 2, and the demultiplexed sequences are also deposited in NCBI PRJNA1085665.

Generation of the initial HASCH SNP set

Identifying the HASCH SNPs set involved three steps: (i) collecting SNP datasets, (ii) filtering and (iii) optimization. Three SNP datasets were used as initial inputs. The WGS7DS are sourced from whole genome sequencing (WGS) data of 383 samples deposited in NCBI and public sites [19–22]. The 21TRICH are sourced from RNA-Seq sequences of 21 trichome samples [24–27]. At the time of data collection (December 2022) these included all *C. sativa* WGS BioProjects with multiple samples, all samples used in *C. sativa* genome assembly projects and all trichome RNA-Seq samples available in NCBI. The sample sequences used and their availability are listed in Supplementary file 3 while the variant calling pipeline used is described in the GATK-Parabricks Benchmarking report [39]. The third set of SNPs is from GBS generated above. The three datasets in vcf file format were filtered to remove the indels and keep the SNPs only using this bcftools pipeline [40],

```
bcftools norm --atom-overlaps . -c w -a --fasta-ref refer·
ence.fasta input.vcf.gz | bcftools filter -i 'TYPE="snp"'
bcftools norm -m +any -Oz -o output.snpsonly.vcf.gz
```

The GBS SNPs vcf file is available at DOI (https://doi.org/10.25918/data.343), while the WGS7DS and 21TRICH SNP matrices can be queried from the ICGRC CannSeek genotype viewer [41]. bcftools was used to merge the three data sets, then SNP statistics were updated using the bcftools fill-tags plugin. Next, only SNPs present in the GBS and (WGS7DS or 21TRICH) datasets and able to pass the following filter criteria were retained: FMISSING<=0.6, MAF is set at >=0.2, and (HWE>=0.0001 or ExcHet>=0.0001). The filtered set was used as input for optimization to identify 2,000 SNPs to be included in the SNP-panel. Integer linear programming (ILP) was used to get the optimal set of markers.

Optimizing HASCH design

The goal for designing the SNP panel was to select subset of markers that provide a maximum of homozygous mismatches between all sample pairs in the input genotype set. The selected markers would further have to be evenly distributed across the genome to be usable for genetic map construction, and each target marker should have minimal SNP variation in its flanking region for effective primer hybridization.

Given initial genotype matrix **G** with **N** samples and **K** markers, encode $\mathbf{g_{i,\,k}}$ as the number of alternate alleles for sample **i** in marker **k**, $g \in \{0,1,2\}$, such that 0: homozygous reference, 2: homozygous alternate, 1: heterozygous, for **i**=1.N, and **k**=1.K. Define 'sample -pair' matrix **P** as an $\frac{N(N-1)}{2}$ by **K** matrix, where $p_{ij,k} = |g_{i,k} - g_{j,k}|_{i<j}$, or the number of allele mismatches between samples **i** and **j** at marker **k**. Missing alleles are encoded in matrix **G** as NA then set to zero in matrix **P** for markers in sample pairs involving them, which means being ignored.

Only homozygous mismatches between pairs were considered in the pairwise mismatch constraints, except in cases where pairs had purely heterozygous mismatches; only for these pairs heterozygous mismatches were considered. So, for each row $\mathbf{p_{ij}}$: change 1s to 0, except if there is no 2 for the row, where 2 means homozygous mismatch, 1 is mismatch between homozygous and heterozygous, and 0 means matching alleles or missing in at least one of the samples. The goal was to identify the subset of a fixed number of SNPs with the maximum number of polymorphisms that could be detected between all pairs in the sample set.

To maximize the number of polymorphisms for a fixed number of marker M from an input set of K markers, the following Integer Linear Programming formulation was defined.

Objective: $\max_x c^T PWX$ maximize number of pairwise mismatches

Subject to:

$\sum X_k \leqslant M$	maximum of M markers to select
$PX \geqslant 1$	at least 1 marker can discriminate any sample pair
$x_k \in \{0,1\}$	binary vector of length K, 1 if include marker k
c	$1-$ vector with length, $\frac{N(N-1)}{2}$

Additional constraints were imposed to evenly distribute the solution across the genome. Aiming at 2000 SNPs over a genome of roughly 1Gb, the genome was divided into equal regions of 500 kb length. For a given region R with D markers from $\mathbf{x_{r+1}}$ ending at $\mathbf{x_{r+D}}$, the constraint is.

$x_{r+1} + x_{r+2} + \cdots + x_{r+D} \geq 1$ at least 1 marker in region R

To minimize the number of SNPs within the flanking region, a marker weight matrix **W** is introduced in the objective function, where W is a K x K diagonal matrix of weights inversely proportional to the number of flanking SNPs within 100 bp in the original unfiltered SNP set. The actual adjustment setting was determined by trial at w=1-0.05$\mathbf{n_{100}}$-0.2$\mathbf{n_{10}}$, where $\mathbf{n_{100}}$ is the number of SNPs within 100 bp, and $\mathbf{n_{10}}$ is within 100 bp having neighbors within 10 bp.

To implement the above optimization problem, a Python script was written using the scipy.sparse module to perform the following:

a. given the vcf file, recode into a {0,1,2} genotype matrix using plink --recode A.
b. divide the genome into regions of equal length, then for each region add the $\sum_{i=r+1}^{r+D} x_i \geq 1$ constraint.
c. generate the weight matrix W from the list of flanking SNPs counts.
d. generate the P matrix from all pairwise combination of samples (row), where $p_{ij,k} = |g_{i,k} - g_{j,k}|$ for all $i < j$. Then for each row, replace all 1's with 0's if there are 2's
e. generate the input LP file following the objective, constraints, and boundary equations in the ILP formulation.
f. use an Integer Linear Programming software to get the optimal solution. We used Gurobi Optimizer [42] with academic license.
g. The solution is the vector of markers selected.
h. Process the solution to get the distribution of mismatching pairs.

Primer design and running of the HASCH platform by a commercial service provider

The 2000 final target SNPs, including 100 bp upstream and downstream flanking sequences were submitted to the commercial service provider Diversity Arrays Technology (DArT) (https://www.diversityarrays.com/) for multiplexed primer design using their proprietary algorithm and design pipeline. After the amplicon selection process 1504 SNP targets (75%) were retained, constituting the final HASCH panel. Running of the HASCH panel, including multiplexed polymerase chain reaction (PCR), amplicon sequencing, de-multiplexing and SNP calling were performed by DArT, using in-house methods as part of their "DArTag" platform.

Multi-dimensional scaling

The WGS dataset was filtered with the 1504 HASCH positions using bcftools. TASSEL5 [43] was used to calculate the distance matrix using identity-by-descent (IBD), and then multi-dimensional scaling MDS with 5 components. The first three principal components (PC) were then plotted in 3d scatter plot using python 'plotly'.

Concordance between HASCH and GBS

The HASCH and GBS vcf files for common samples genotyped by both platforms (Supplementary file 4) were filtered to include only common samples and sites, and samples to have minimum 90% call rate using bcftools [40]. Bcftools -stats calculates the per-site discordance PSD, and genotype concordance by sample SNPs GCsS. From the documentation, non-reference discordance NRD.

$$\mathrm{NRD} = 100 * (\mathrm{xRR} + \mathrm{xRA} + \mathrm{xAA}) / (\mathrm{xRR} + \mathrm{xRA} + \mathrm{xAA} + \mathrm{mRA} + \mathrm{mAA})$$

where x means mismatches, m means matches. AA, RR, and RA are homozygous alternate, homozygous reference, and heterozygous respectively. Simple discordance metric can be over optimistic due to typically large mRR. In NRD, mRR is excluded from the denominator [44]. Simple concordance C is calculated,

$$NRD = 100 * (xRR + xRA + xAA) / (xRR + xRA + xAA + mRA + mAA)$$

The site and sample C and NRD distributions are plotted using python matplotlib.

Comparative phylogenies

HASCH and GBS datasets were filtered to use the same sample of one replicate from the IPK samples (Supplementary file 5). HASCH sites were filtered to MAF above 0.2 and missing rate below 0.1. GBS sites were filtered

to MAF above 0.3 and missing rate below 0.01. Each phylogenetic tree was calculated by Neighbour-Joining using TASSEL5. The resulting Newick files were compared using VisualizeMatching SharedPhylogeneticInfo from the TreeDist R package to draw two trees side-by-side with highlighted common branches. Branches sharing the same set of nodes were checked manually and assigned similar color. The distance matrices by IBD were calculated by TASSEL5. All sample pairwise distances from HASCH were plotted in the y-axis, vs. from GBS in the x-axis in a scatter plot.

Comparative heterozygosity rates

Using the same HASCH and GBS datasets to generate the comparative phylogenies, the sample heterozygosity were calculated for each set using bcftools stats. The Per-sample Counts (PSC) rows give the number of homozygous references, homozygous alternate and heterozygous. The percent heterozygosity was then calculated for each sample, and their distribution plotted using python matplotlib.

Mismatch SNPs in hemp HASCH and medicinal WGS7DS

Using only one replicate per sample and excluding the Inbred (Fn) samples from the HASCH dataset (Supplementary file 6), the number of mismatches between homozygous sites were counted for all sample pairwise comparisons. Similarly, homozygous mismatch count distribution was plotted for drug type or type II samples from the WGS7DS (Supplementary file 7). The distribution of mismatch counts on all pairs was plotted using python matplotlib.

CBD quantification using GC-MS analysis

Analysis of CBD content was performed by GC-FID Chromatography using Agilent 6850 A gas chromatograph with Flame Ionization Detection (FID), with a BPX5 model no- SGE 054117 MS column (50 m x 200 µm x 1 µm) (Phenomenex, Torrance, California, United States). Samples were directly injected from the auto sampler held at 50˚C for 1 min then increased to 300 ˚C at 8 ˚C /min and held for 10 min making the total run time per sample 42.25 min. Hydrogen was used as the GC carrier gas at a flow rate of @1.2 mL/min. Hydrogen at a flow rate of @30 mL/min and compressed air at a flow rate of @300 mL/min were used as the combustion gases. A CBD standard (Sigma Aldrich) was used for peak identification and generation of calibration curves for quantitation. Data were recorded and processed using Openlab Software (Version 3.6).

Genetic map and QTL mapping for CBD content

Genetic map and QTL mapping used combination of R/qtl [45] and ASMap packages [46]. HASCH genotyping data of the F2 population was diagnosed for low call rate, as well as pairs of unusually similar genotype data in R/qtl. Monomorphic markers and markers with sample genotyping call rates of less than 90% were filtered out. The resulting genotype input was further filtered in ASMap for several parameters including segregation distortion, identical genotype data, evidence for genotyping error [47] and markers with duplicated information due to co-location on the same position. The linkage map construction was conducted using R/qtl. The threshold for placing two markers in same linkage group used an estimated recombination fraction maximum of 0.35 and minimum LOD score of 0.6. Inter-marker distances in centiMorgans (cM) were estimated using the kosambi function. Re-estimation of the final genetic map used Lander-Green algorithm [48] employed in est.map function of R/qtl. The genetic map was utilised for QTL mapping for cannabinoid content using log-transformed percent values in F2 population comprised of a set of 121 individuals with detectable CBD contents. Single QTL model using Haley-Knott regression, two-dimensional scan, and multiple QTL analyses were carried out. Single-QTL analysis used a density of 1 cM while the two-dimensional scan used 2 cM. The final QTL model was obtained from "stepwiseqtl' analysis. The 95% Bayes credible c intervals around the maximum likelihood estimate of the QTL location was estimated using the "bayesint" function. The proportion of phenotypic variance explained by the QTL was estimated using "fitqtl" function. The genetic and QTL map was drawn using MapChart [49].

Results

HASCH platform design

SNP calling using the three datasets (WGS7DS, 21TRICH and in-house GBS) against the CBDRx reference genome resulted in a total of 92,205,134 SNP, with 90,601,645 (94.2%) from WGS, 5,408,878 (5.6%) from the 21TRICH RNAseq and 194,611 (0.2%) from in-house GBS (Fig. 1a).

The merger of all in-house GBS SNPs with intersections of WGS7DS and 21TRICH resulted in an initial SNP set of 194,611 SNP (grey area in Fig. 1a). Filtering for 0.6 call rate (Fig. 1b) and 0.2 minor allele frequency (MAF) (Fig. 1c) yielded a filtered set of 57,251 SNPs with 31,781 SNP (55,5%) common to all three data sets and 25,470 (44.5%) common to WGS7DS and in-house GBS (Fig. 1d).

Integer Linear Programming (Fig. 1e) was used to select 2,000 target SNPs from the filtered set that fulfilled the constraints of (i) at least 1 SNP per 500 kb, (ii) maximum number of polymorphic homozygous SNPs in pairwise comparisons between all sample entries ($N=590$) and (iii)

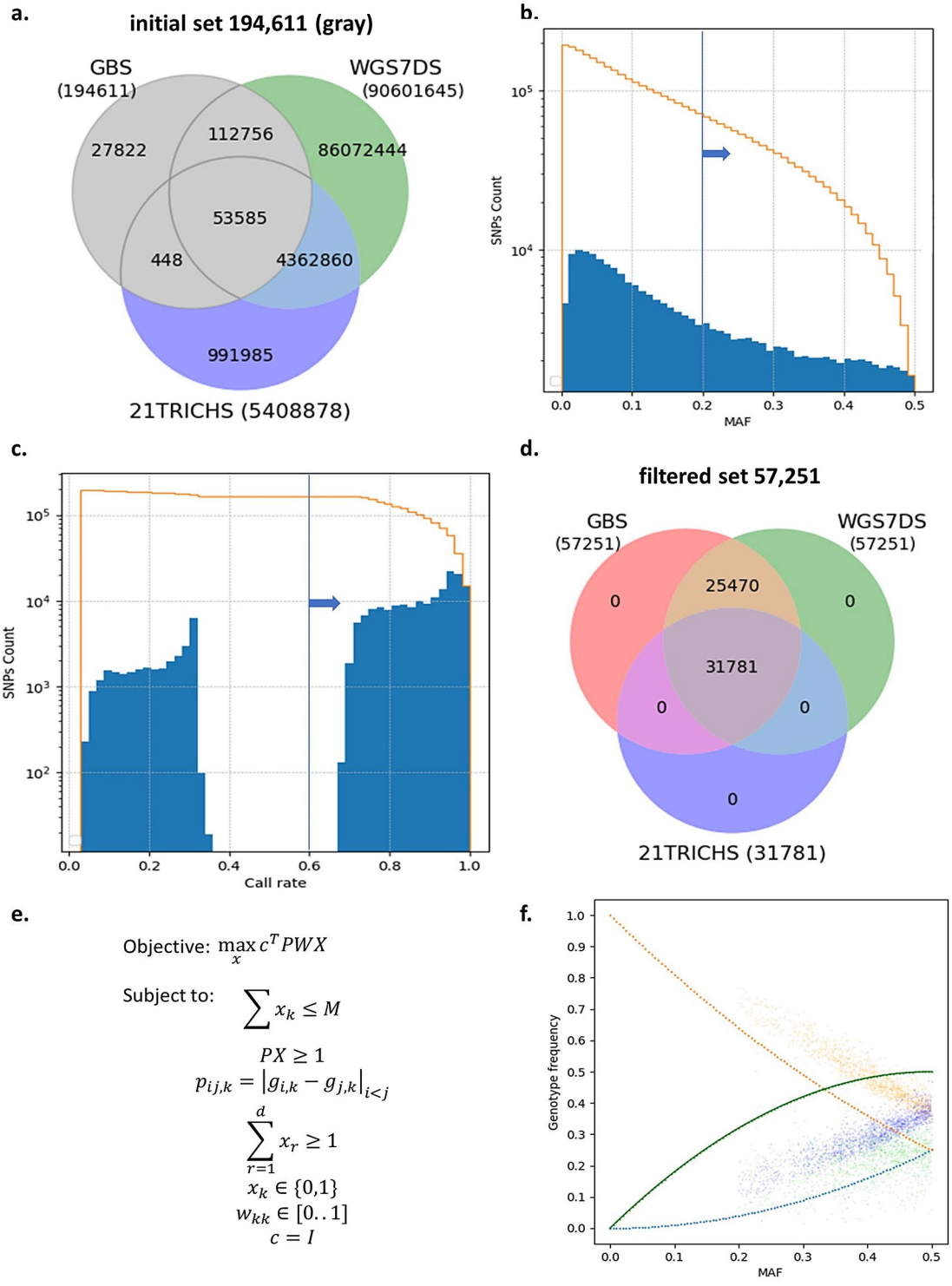

Fig. 1 HASCH design. The initial SNP set comprised **(Grey area in a)** the merger of in-house GBS, with their intersections with public WGS (WGS7DS) and RNAseq (21TRICH) data. The **(b)** Call Rate and **(c)** Minor Allele Frequency (MAF) distribution were plotted and a cut off of at 0.6 Call Rate and 0.2 MAF (blue lines and arrows) were used as a filter. **(d)** The filtered set consisted of 57,251 SNPs, with 25,470 common to GBS and WGS, and 31,781 common to all three datasets. **(e)** Integer Linear Programming was used to select 2,000 SNPs from the filtered set, maximizing the number of pairwise polymorphisms. **(f)** After primer design, 1,504 targets were retained, and their MAF plotted against genotype frequencies, with Hardy-Weinberg proportions superimposed as lines. Orange = aa (q^2). Green = Aa (2pg), Blue = AA (p^2)

minimize number of SNP in 100 bp regions flanking target SNPs.

After the proprietary amplicon selection process from Diversity Array Technologies (DArT), 1504 targets (75%) were retained (Supplementary file 8). Of these 1,289 (86%) were common to GBS, WGS, and RNAseq and 215 (14%) common to GBS and WGS. Plotting genotype frequency by MAF revealed a near Hervey Weinberg distribution skewed towards high minor allele frequencies (MAF) and low heterozygous (2pq – green) rates (Fig. 1f).

Of the 1504 SNPs, 1,069 markers (71%) localized within coding region of annotated CBDRx gene models, and 435 (29%) were located within intergenic regions (Supplementary file 9). Transition SNPs comprised 63% of the SNPs, including 476 A/G and 476 C/T. Transversions SNPs made up 37% of the total polymorphisms with 122 A/C, 115 G/T, 204 A/T, and 111 C/G.

Autosomal chromosomes had between 110 and 194 markers per chromosome, while the X-chromosome contained 203 markers (Fig. 2a, Supplementary file 8). Three markers located on scaffolds not assigned to CBDRx pseudochromosomes. The average physical distance between two adjacent SNPs across the whole genome in the HASCH set was 566 kb (Fig. 2b). More than 50% of the markers are spaced from each other at a distance of 474 kb or less.

Type I = THC dominant; Type II = THC-CBD balanced; Type III = CBD dominant, Hemp types = CBD dominant, Drug types = THC dominant, NLD = Narrow leaf drug type; BLDT = Broad leaf drug type.

To test the utility of the final HASCH in discriminating within a broad germplasm set we extracted the respective HASCH SNP data (N = 1504) from the publicly available WGS7DS data on 383 accessions and constructed a multi-dimensional scaling (MDS) plot across the first three principal components (PC) (Fig. 3). Entries were color-coded using the annotations of Cannabis types provided by the data owners. PC1 broadly separated low-THC Hemp-types (Type III, Hemp-type and Hemp), including "Basal Cannabis", from high-THC Drug-types (Type I, Drug-type, Narrow leaf Drug type (NLDT), Broad Leaf Drug type (BLDT) and Feral Drug types). Type II accessions of intermediate THC and CBD content were found dispersed across both the Hemp-type and the Drug-type clusters. PC 2 separated Basal Cannabis from Hemp-types. PC2 and PC3 distinguished well within Hemp-types and Drug-types. Unknown samples scattered throughout the plot and were easily associated with certain Cannabis types based on their cluster proximity.

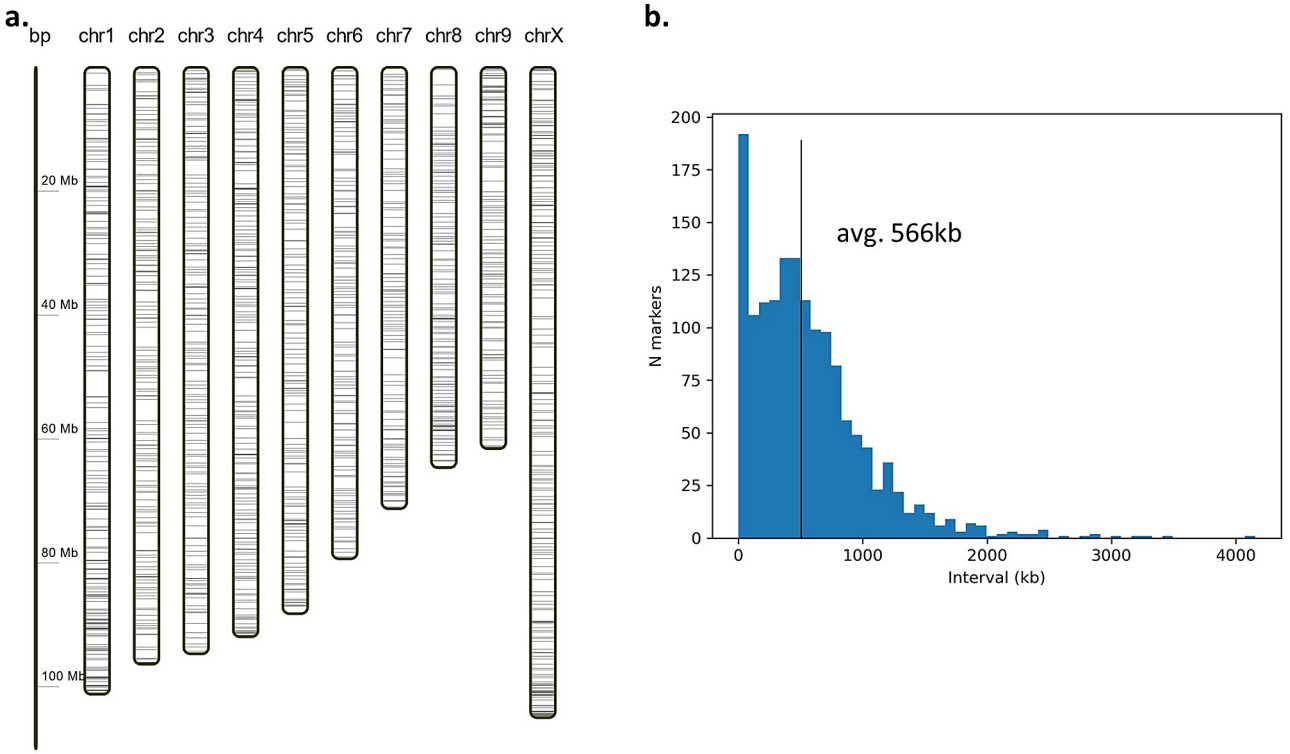

Fig. 2 HASCH SNP distribution. **(a)** Genome-wide distribution of the final 1,504 target SNPs across the CBDRx reference, and **(b)** corresponding distribution of intervals between adjacent SNPs

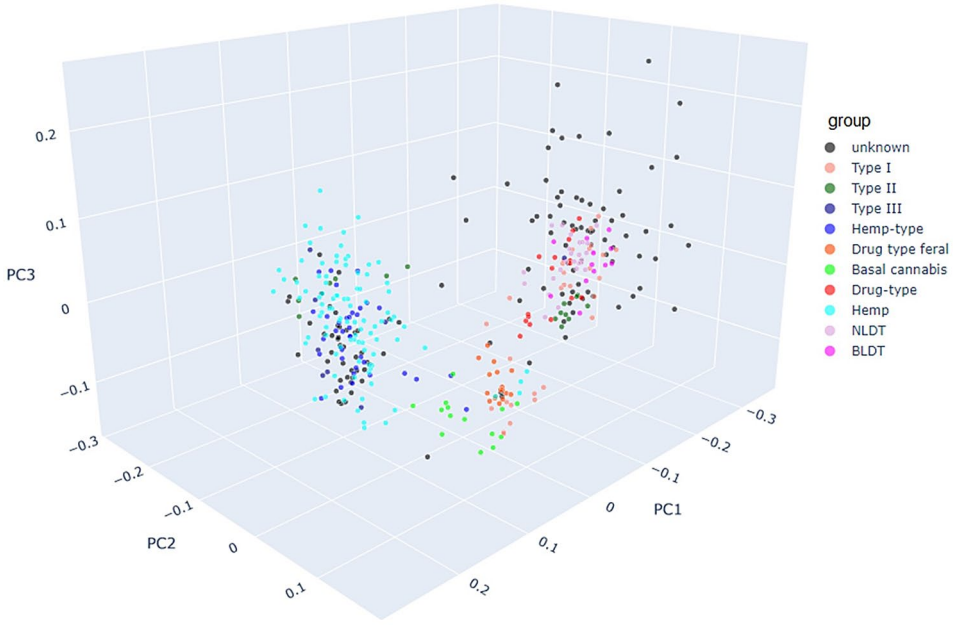

Fig. 3 In silico analysis of HASCH utility to discriminate diverse Cannabis germplasm. Multi-dimensional scaling plot showing the first three principal components (PC) using the ,1504 target SNP of the HASCH taken from publicly available genotype data from diverse accessions with respective chemotype classifications (*N* = 383). Blue colorations indicate Hemp types (dark blue = Type III, blue = Hemp-type, turquoise = hemp); red colorations indicate drug types (salmon = Type I, red = Drug type, orange = Drug type feral, light pink = NLDT (Narrow leaf drug type), dark pink = BLDT (Broad leaf drug type); Green colorations indicate intermediate and basal types (dark green = Type II; green = Basal Cannabis); dark grey coloration indicates "unknown". Type I = THC dominant; Type II = THC-CBD balanced; Type III = CBD dominant, Hemp types = CBD dominant, Drug types = THC dominant, NLD = Narrow leaf drug type; BLDT = Broad leaf drug type

HASCH platform validation

SNP call rates and heterozygosity were empirically determined in the HASCH using 376 independent DNA samples (Supplementary file 10). The mean call rate across all SNPs was 97% (Supplementary file 11), and the mean heterozygosity observed among SNPs was of 25%.

Genotypic concordance between HASCH outputs and GBS outputs was estimated and averaged across overlapping SNPs in 163 commonly genotyped samples (Fig. 4). Filtering for 0.9 Call rate in HASCH and GBS resulted in a set of 1,385 overlapping SNPs. Concordance and Non-Reference Discordance (NRD) [44] values across SNPs (*N* = 1,385) averaged at 0.916 (Fig. 4a) and 0.116 (Fig. 4b) respectively. Average concordance across samples (*N* = 163) was 0.925 (Fig. 4c) while average NRD across samples was 0.097 (Fig. 4d).

To compare utility for phylogenetic studies, the GBS data for one replicate each of the 55 IPK accessions were filtered for MAF above 0.3 and call rate of 1.0, resulting in 5,582 SNP. HASCH data for the same 55 accession and replicate was filtered for 0.9 call rates and 0.2 MAF resulting in 1,252 SNPs. Phylogenetic trees obtained with both SNP set were near identical (Fig. 5a) and genetic distances based on identity by descent (IBD) where highly correlated (Fig. 5b) with the squared Spearman's rank correlation (r^2_sp) at 0.833 and the squared Pearson's correlation coefficient (r^2_p) at 0.829.

To test the utility of HASCH in determining levels heterozygosity we compared the filtered HASCH and GBS genotyping datasets for the same samples. While the average percentage of heterozygosity in the 55 lines for the 1,252 HASCH SNPs was 31.5% (Supplementary Fig. 1a), it was 40.1% for the 5,582 GBS SNPs (Supplementary Fig. 1b).

In addition, we genotyped a small set of lines that had been successively inbred from F1 to F4 (Supplementary Fig. 2), demonstrating that HASCH was able to monitor reducing levels of heterozygosity through successive rounds of inbreeding from around 30% at F1 to around less than 10% at F4.

HASCH genotyping information on 87 diverse Hemp-type accession were used in pairwise comparisons to assess the number of polymorphic SNPs between any two accessions. The average polymorphic SNPs across all pairwise combinations (*N* = 3,741) was 272 (Fig. 6a).

Similarly, HASCH information on 116 diverse Drug-type accessions filtered from the WGS7DS dataset revealed an average number of polymorphic SNP in pairwise comparisons (*N* = 6670) of 251 (Fig. 6b).

Genetic map and QTL mapping using HASCH

The utility of the HASCH for the construction of a genetic map was determined using HASCH data on a feminized F2 population derived from a cross of a

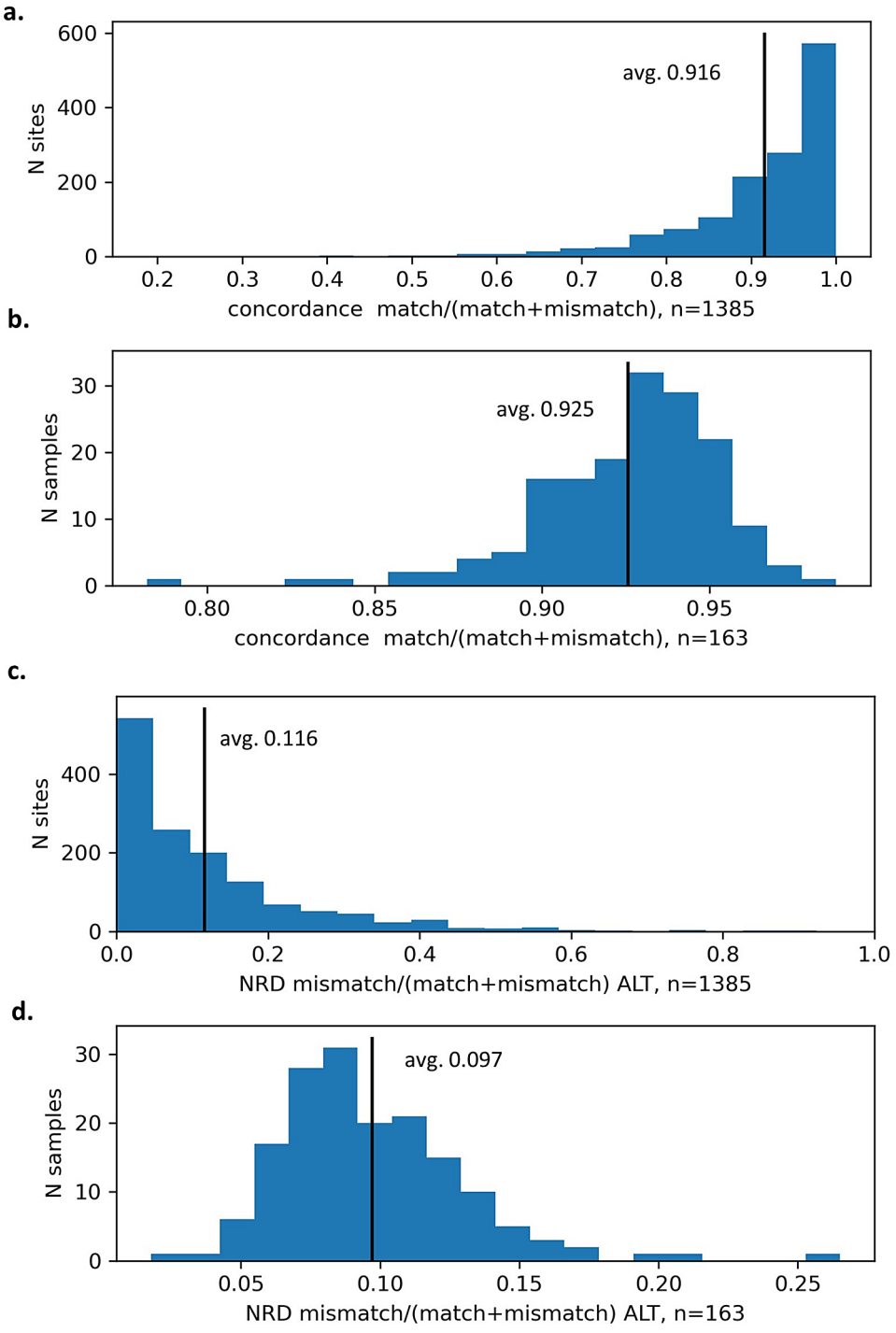

Fig. 4 Concordance and non-reference discordance (NRD) between GBS and HASCH of samples that were genotyped with both platforms. Concordance by **(a)** SNP site (*N* = 1385) and **(b)** Samples (*N* = 163) as well as non-reference discordance (NRD) by **(c)** SNP site (*N* = 1385) and **(d)** Samples (*N* = 163) for common SNPs filtered to above 0.9 Call Rate in the HASCH data. bcftools stats was used to get the Per Site Discordance (PSD), and Genotype concordance by sample (GCsS). Black lines indicate averages (avg.)

CBD-dominant line (IPK_57) and a THC-dominant line (IPK_36). From the 153 F2 samples, one sample had no genotype call in all markers. Sample diagnosis resulted to dropping five samples identified to have 80–100% matching genotype data with other sample, as well as one sample with no genotype call in all markers.

Filtering for monomorphic markers and markers with less than 90% sample call rate, resulted in 647 SNPs, which were further reduced to 313 SNPs by filtering in

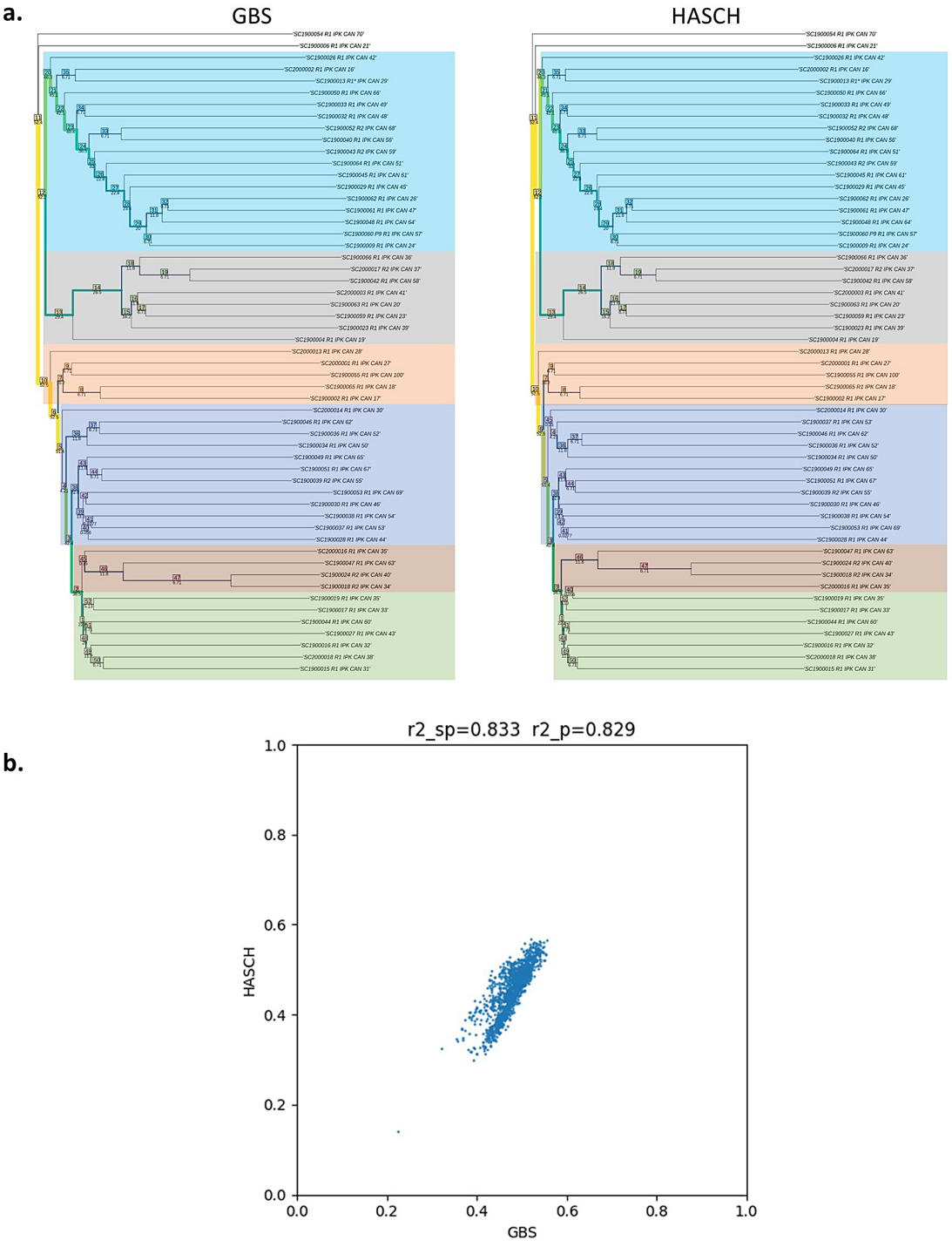

Fig. 5 Comparison of HASCH with matching GBS genotyping data of the same sample for each IPK accession for phylogenetic and genetic distance analysis. **(a)** Phylogenetic tree comparison using filtered GBS (MAF > 0.3, no missing call, 5582 SNPs) and HASCH (MAF > 0.2, call rate > 0.9, 1252 SNPs). Branches with the same set of nodes are coloured the same. **(b)** Corresponding correlations (r2_sp = Spearmans, r2_p: Pearsons) of genetic distances (Identity by Descent) between the GBS and HASCH results

preparation for linkage map construction in ASMap (Supplementary file 12). These 313 SNPs were anchored to 12 linkage groups in linkage map construction in R/qtl [45]. Two of these linkage groups were composed of two, and one unlinked marker respectively, and were later dropped in the map. The remaining ten linkage groups corresponded to nine autosomes and the X chromosome of the *C. sativa* genome. Based on marker naming, which carried the chromosome and physical position, no marker was misplaced on another chromosome. The plot

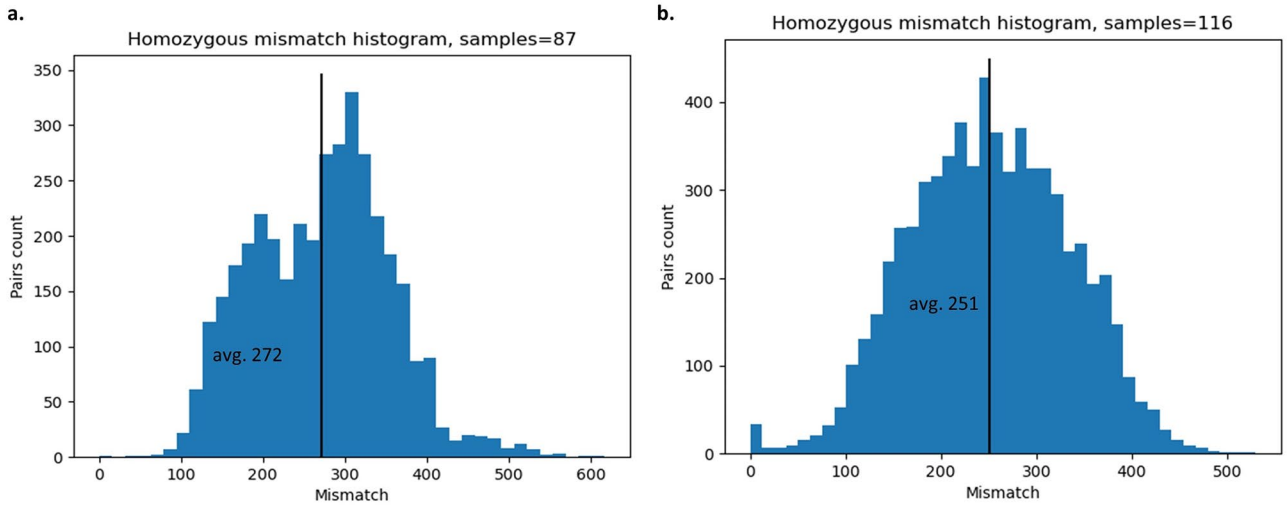

Fig. 6 Marker utility in pairwise sample comparisons. Pairwise homozygous SNP mismatches within **(a)** industrial Hemp Diverse set (*N* = 87) based on HASCH genotyping data and **(b)** medicinal Cannabis diverse set (*N* = 116) based on HASCH training data. For N samples, the number of homozygous mismatches between every N(N-1)/2 sample pair combination were counted. Black lines indicate averages

linkage map was arranged based on chromosome number (Fig. 7a). The linkage map covered a total of 582.7 cM of the genome. Chromosome 6 with 73 cM was the longest, while chromosome 7 with 37.3 cM was the shortest. The mean inter-marker distance across all chromosomes was 1.93 cM, with the largest inter-marker distance of 26.9 cM observed in chromosome X.

Utilising this linkage map, a single putative QTL for cannabidiol content, *qCBD7*, was consistently discovered in chromosome 7 in all methods explored in our QTL analysis. This QTL had an LOD score of 21.9, well above the threshold based on 1000 permutation test at 5% significance level, and explained 57% of the phenotypic variance for the cannabidiol content in the F2 population. The Bayes credible confidence QTL interval estimate was located between 21.3 and 24.7 cM of chromosome 7, with maximum likelihood estimate at 22.5 cM. Scrutinizing markers located at the short genetic distance covering the QTL confidence interval revealed three peak markers at 22–38 Mb physical position (Fig. 7b), as well as one peak marker at 63 Mb. The phenotypic values for the allele states of the peak marker at 63 Mb demonstrated a good marker trait association (Fig. 7c).

Discussion

While genotyping technologies and software can now identify millions of genetic markers, there is need to reduce marker densities to be manageable, informative and cost-efficient for routine applications. The quality, informativeness and utility of fixed SNP sets largely depends on input data selection and SNP selection methods from the input data.

For input data selection we decided on a combination of publicly available and de-novo generated raw datasets called in-house against the CBDRx reference (Fig. 1). The bulk of input data was low coverage WGS7DS data (65% of samples and 94.2% of SNPs). The 21TRICH RNAseq data (4% of samples and 5.4% of SNPs) and in-house GBS data (31% of samples and 0.2% of SNPs) of higher read depth served to build confidence into the SNP calls. The incorporation of RNAseq expression data further anchored the majority of SNPs (71%) in existing gene models. This was intentional to reduce the number of heterozygous calls and increase the number of polymorphic SNPs in pairwise comparisons. SNPs in coding regions are subject to selection pressure and thus are more likely to get fixed through either natural or artificial selection [50]. The in-house GBS data further served as comparator to the HASCH data generated on the same DNA samples for downstream validation (Figs. 4 and 5).

Feature selection methods have been applied to optimize marker panels for specific end uses and can be classified into filters, wrappers or embedded methods [51]. The traditional and fastest approach is the filtering method. Each marker has some statistical or biological attributes, and filtering sets a cut-off for continuous, or specific values for categorical properties. The selection is independent for each marker. For SNPs the typical filtering criteria include Minor Allele Frequency (MAF), heterozygosity rates, read depth, distribution across the genome, local SNP density and Linkage Disequilibrium (LD). Wrapper methods create models that consider only a subset of markers as inputs to fit with training datasets, with the objective of finding the smallest subset that best fits the training set. Wrapper methods traditionally use statistical linear models, for example Best Linear Unbiased Prediction (BLUP) in genomic selection. Heuristic models utilizing machine learning approaches

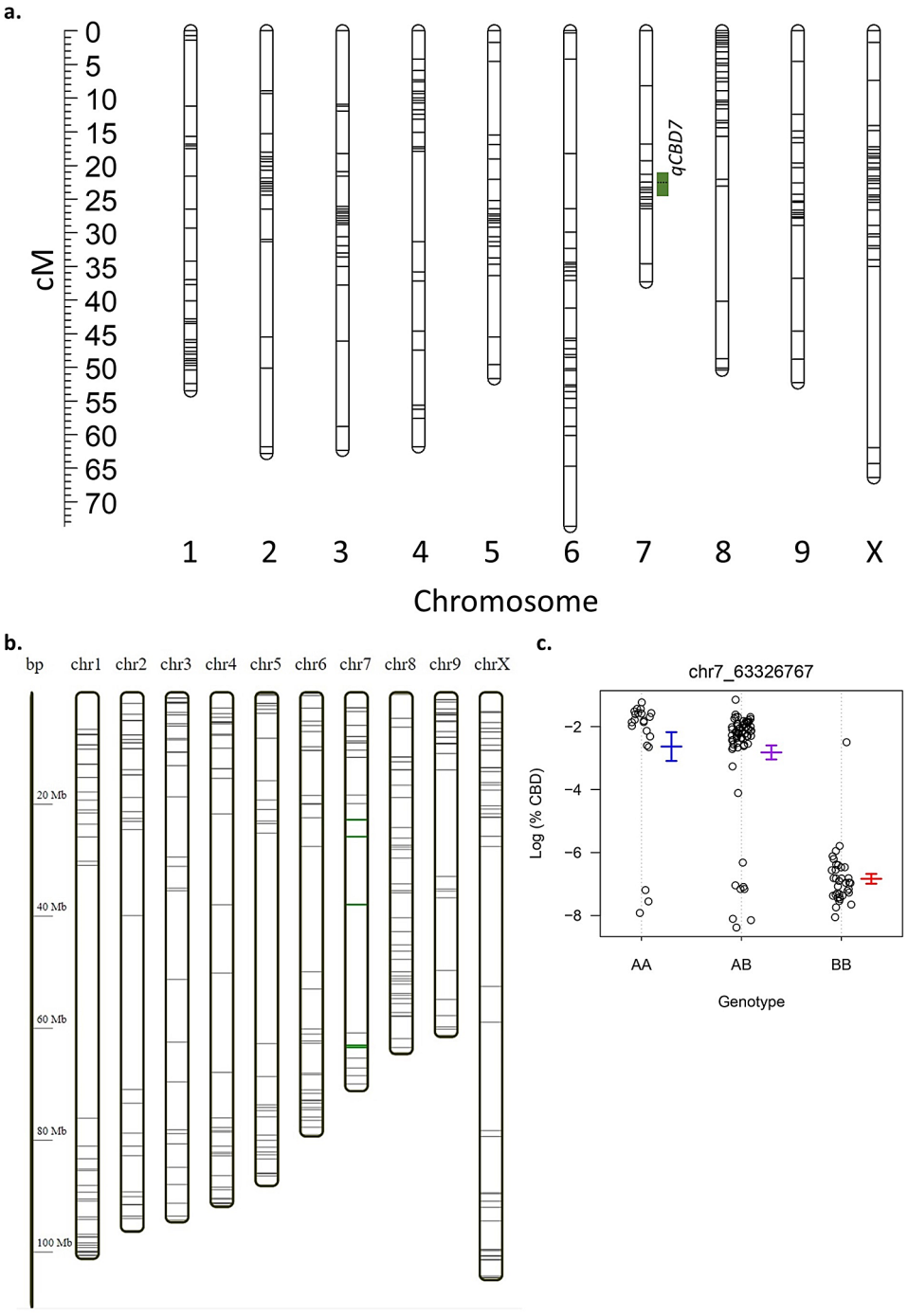

Fig. 7 Genetic map and QTL mapping for CBD content in a F2 mapping population. **(a)** Genetic map of a segregating F2 population derived from an THC-dominant IPK_CAN_36 (Male) and CBD-dominant IPK_CAN_57 (Female) cross based on 313 markers. **(b)** The genetic markers in their physical location in the genome. **(c)** The genetic position of qCBD content for F2 lines of different peak marker alleles at 62 Mb (*N* = 212)

and artificial intelligence, specifically decision trees and genetic algorithms have been used recently [52, 53]. Embedded methods combine filters to reduce input size, and wrapper methods for accuracy with respect to the training set. The training set are usually phenotypes of interest and must be available for the samples used in the genotype dataset.

While there are many potential applications from heuristic methods to address SNP selection challenges, employing optimization based on mathematical programming remains limited. Most notable is the identification of haplotype blocks and selection of tag SNPs using software like Haploview and Tagger [54–57]. Informative SNPs were also selected for paternity inference

using integer linear programming (ILP) by Nishiyama [58]. ILP has wide application to computational biology [59], but to our best knowledge has not been utilized to solve marker selection problems in genotyping platform design.

Subset selection from the discrete set of markers to optimize an objective function is a combinatorial problem. Recent algorithms from Machine Learning approaches such as genetic algorithms and decision trees can be used on these problems, however they need heuristics and implementation-specific settings to represent specific biological requirements. ILP enabled formulation of our requirements into simple linear equations. Our main goal was to identify a minimal set of SNPs that could differentiate all samples from WGS, RNAseq and in-house GBS data. For design purposes, we fixed the number of markers to be at most M=2000, $\sum x \leq M$, and formulated a related problem of maximizing the count of polymorphisms, $c^t PWx$, across all sample pairs. Additional considerations to satisfy included: (i) all sample pairs must be distinguishable by a subset of markers in the solution $PX \geq 1$; (ii) for application to QTL mapping the markers must be evenly distributed across all chromosomes; (iii) the presence of variants in proximity to target variants should be minimal as they can interfere in primer design and hybridization, represented with diagonal matrix $\mathbf{W}$ with weight $\mathbf{w_{kk}}$ for marker k.

We used our ILP approach on a filtered set of 57,251 SNP (Fig. 1) to determine a subset of 2,000 SNPs that maximizes homozygous mismatches. After primer design, the resulting final 1504 SNP had an even coverage of the genome with an average of 567 kb between two adjacent SNP and a minimum of 110 SNPs per chromosome (Fig. 2) at an average MAF of 0.32. Uniform SNP distribution in combination with high MAF maximizes utility for most breeding and pre-breeding applications. In QTL mapping this increases the number of polymorphic SNPs between any two potential parents and enhances the chances to detect recombination events while minimizing the distance of peak markers to the causal variation. In phylogeny, equal distribution and high MAF ensure that a diverse set of genetic markers is available, enhancing the resolution and accuracy of phylogenetic trees especially if certain regions are overrepresented or underrepresented. Rare variants may have limited impact on association studies, so a higher MAF improves statistical power in association or genomic selection studies. Twenty-three gaps of over 2 Mb were found after filtering with the proprietary primer design (Fig. 2, Supplementary data 8). There is an opportunity to fill these gaps in genome coverage through supplementation. Unlike fixed arrays, the amplicon-based design is flexible, allowing for inclusion of additional markers to fill gaps, replace markers of low call rate and to add value through addition of specific trait markers as they become available to the public domain. The designed density and even distribution was seen as suitable for biparental QTL mapping studies, which was further empirically tested. Although the HASCH primer design and running in production were outsourced to DArT, the 2000 SNP targets we provide (Supplementary file 13) can be utilized to design in-house amplicon panels. Open source primer design software and protocols for highly multiplexed PCR like PrimerMapper, V-primer or Ultiplex, based on Primer3 [60–63], are available and should provide workable solutions. In HASCH primer design 496 of the original 2000 targets (25%) were excluded. While it is likely that different primer selection approaches would result in slightly different final sets, it is not expected overall performance for the demonstrated use cases would be affected.

To test the utility and accuracy of the HASCH, results were compared to GBS results acquired for the same samples. Concordance was high, in respect to both SNP (Fig. 4a) and samples (Fig. 4b) indicating that the HASCH reliably calls SNPs, including heterozygous alleles. However, at 92% average SNP concordance was lower than reported for a comparable amplicon panel in rice (1k-RicA) where rates of 99% were reported [64]. This is likely due to the fact that inbreeding rice has significantly lower rates of heterozygosity, then outcrossing *C. sativa*. In the respective studies rice had an average heterozygosity rate of 1.5%, while for *C. sativa* it was 25%. Heterozygous calls are more difficult to ascertain than homozygous calls. In addition, the rice study was able to draw on a higher abundance of high-quality data for design. In terms of call rate, HASCH at 97% and 1k-RiCA with an average call rate of 95% were comparable.

A HASCH-based MDS plot, filtered from the WGS7DS data, was able to discriminate between and within drug- and hemp-types (Fig. 3) suggesting broad utility across the genetic diversity of the *C. sativa* gene pool.

To demonstrate the versatility of HASCH in genetics and pre-breeding applications a number of case studies were investigated. Comparison between phylogenetic trees constructed either by heavily filtered GBS SNP data (N=5,582) or less stringent filtered HASCH SNP data (N=1,252) with 404 overlapping SNP targets, showed high levels of congruence in tree structure (Fig. 5a). This finding was further supported by high levels of correlation in genetic distances between entries calculated using either data set (Fig. 5b). However, when compared to untargeted GBS, HASCH seemed to underrepresent the true rates of heterozygosity (Supplementary Fig. 1). This type of ascertainment bias (or sampling bias) was expected due to HASCH design criteria (Fig. 1), which selected for fixed SNPs that maximize pairwise homozygous mismatches within the available dataset. However,

when tested on successively inbreed cultivars HASCH was able to detect a reduction of heterozygous loci for each round of inbreeding.

The HASCH was specifically designed to have high levels of homozygous polymorphic SNP between any two parents (Fig. 6). While the average number of polymorphic SNPS between any two industrial Hemp or medicinal Cannabis accession was only in the order of around 250 SNPs (Fig. 6) the number is likely to be higher in actual crosses, where parents tend to be genetically more distant.

Parents of the F2 population in our proof of concept study shared a total of 647 polymorphic SNPs. The 313 segregating markers comprised a genetic map of 10 linkage groups, which is consistent with published de-novo assemblies [18] and genetic maps [18, 29, 31]. With a total size of 582.7 cM the genetic map was considerably smaller than that of published *Cannabis sativa* genetic maps based on WGS/GBS data. Using 1,235 total segregating markers Grassa et al. [18] reported a linkage map consisting of 10 linkage groups with a map distance of 818.6 cM and a mean inter-marker distance of 0.66 cM. Woods et al. [29] used 1,817 markers in the genetic map that identified 10 linkage groups. While they did not report the total distance of their genetic map, their Fig. 2 suggests to be >1000 cM. While typically a shorter map length map is preferred [65], differences in map lengths could be attributed to difference in the chromosomal recombination frequency that are specific to each mapping population used, the genetic distance between the parents, size of the population, as well as marker type and density used [66].

QTL mapping using HASCH platform for total CBD content reliably detected qCBD7, a major QTL on Chromosome 7 with a short genetic distance (3.4 cM) of confidence interval with very high phenotypic variance explained (Fig. 7a). Three peak markers in our QTL confidence intervals had physical positions at 22-38 Mb, overlapping with the genomic region at 26-31 Mb harbouring 13 cannabinoid synthase homologs in the CBDRx genome reported by Grassa et al. [18], demonstrating the utility for HASCH in QTL mapping and detection. An additional peak marker is at 63 Mb showed good marker trait association (Fig. 7b), which merits further investigation.

Collectively, comparison between GBS and HASCH performance and a number of genetic case studies suggest HASCH to be of high utility in applied *C. sativa* research and development. With 1504 genome-wide target SNPs HASCH sits at the lower end of mid-density genotyping platforms and fills a gap that is not covered by untargeted approaches such as GBS or higher density arrays that range in the order of several 10,000 SNPs and are typically one order of magnitude more expensive [67, 68]. Conversely running multiple single target assays (e.g. KASP or PACE markers) is not cost efficient anymore above a maximum of ~50 targets [6, 69]. Amplicon based SNP platforms in the range of 100–2000 SNPs have been demonstrated as versatile tools in pre-breeding and breeding applications for a number of crops including rice [64], buckwheat [70] and Japanese cedar [71]. At 384-plexing the HASCH platform is suitable to run two QTL mapping populations per batch. As opposed to untargeted approaches such as GBS, fixed SNP sets has the advantage of generating a complete SNP matrix of the same positional composition every run with minimal processing of raw data required. This makes amplicon panels accessible to researchers with minimal computational infrastructure and capacity. Furthermore, it greatly facilitates downstream analyses and enables meta analyses between experiments and/or populations. These features make HASCH attractive for breeding applications such as genomic prediction. While we did not test for HASCH utility in this respect, the 1kRiCA for rice [64], a SNP amplicon panel of similar design with an average physical marker interval of 372 kb, was demonstrated to predict performance for major agronomic traits with acceptable accuracies. Thus, investigating HASCH, with an average interval of 566 kb, for suitability in genomic prediction applications, such as the generation of genomic estimated breeding values (GEBV), is merited. Low cost per sample and fast turn-around time of genotyping solutions are crucial in a breeding context, where costs per sample constrain population sizes and in-season data analysis is required to make selection decisions for the next cycle. In our experience, running HASCH over GBS reduces genotyping cost by 5–8 times per sample, at a turnaround of 2–3 weeks compared to 2–3 months for GBS. The script to generate the input file for ILP optimization from vcf file is available at DOI (https://doi.org/10.5281/zenodo.11149359). It is expected to be usable for any vcf file regardless of species, ploidy, chromosome number, or nomenclature, and is limited only by the information available in the vcf and plink files. Heterozygosity affects the solution by limiting the marker choices to avoid potential interference during primer hybridization. They are ignored in mismatch count except when essential, and could be revised to penalize their presence thus favoring highly homozygous targets, if desired.

Taken together our ILP approach resulted in HASCH, a genotyping solution of high utility for selected use cases and potential for other breeding applications. We expect ILP will serve as a valuable tool in the development of custom genotyping platforms for other underdeveloped plant and animal species.

Conclusion

Our study demonstrates the feasibility of Integer Linear Programming (ILP) to design a mid-density genotyping panel from heterogenous source data. In the absence of readily available high-confidence targets, and provided the design criteria and objectives can be expressed as linear equations, ILP is more explainable than machine learning and less biased than filtering methods. This makes ILP an attractive approach for the design of genotyping solutions for other species of limited public domain NGS data. Further evaluation using more constraints or weights is recommended as more information about the markers or samples becomes available.

HASCH is the first publicly available mid-density genotyping platform for *C. sativa*. Its design specifications and validation results across a broad range of distantly related cultivars suggest robust utility for a variety of applications in medicinal cannabis and industrial hemp research and development, particularly in respect to genetic resource management and (pre-) breeding. Through a number of phylogenetic and quantitative genetic case studies we demonstrated that HASCH performs comparable to high-density untargeted genotyping platforms, while being cheaper and faster and requiring minimal bioinformatics capacity for data management and analysis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10734-z.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The authors would like to acknowledge Andrew Kavasilas for enabling this work as industry partner of LP210200606 and for ensuring that it aligns with industry needs. We acknowledge the Australian Research Council (ARC) for funding under LP210200606. We further acknowledge the work of Diversity Arrays Technology (DArT) in particular Andrzej Kilian for helping with the final selection of target SNPs and for hosting HASCH as a service provider. Finally, the authors acknowledge the provision of computing and data resources provided by the Australian BioCommons Leadership Share (ABLeS) pro-gram, which is co-funded by Bioplatforms Australia (enabled by NCRIS), the National Computational Infrastructure and Pawsey Supercomputing Centre.

Author contributions

TK and RM conceived and conceptualized the study and supervised LM. LM designed the work and carried out most bioinformatic analyses. ET and RD carried out the QTL mapping study including establishment of the population, phenotype data generation and data analysis. JM and AB carried out the GBS data acquisition. LGD carried out the in-breeding experiments.

Funding

This study was funded by the Australian Research Council (ARC) Linkage project LP210200606. In addition, first author Locedie Mansueto received a stipend from Southern Cross University (SCU).

Data availability

The datasets generated and/or analysed during the current study are included in this pub-lished article and the supplementary information files. The WGS7DS SNP matrix was generated using the Parabricks Genomic sequence variant-calling, while the 21TRICH matrix using the RNA-Seq sequence variant-calling pipelines both available at (https://doi.org/10.5281/zenodo.10685744) using public sequences listed in Supplemen-tary file 3. The GBS matrix was generated using the TASSEL-GBS pipeline using the demultiplexed sequences listed in Supplementary file 2 submitted to NCBI under project PRJNA1085665. The generated SNPs vcf file from GBS is available as DOI (https://doi.org/10.25918/data.343), while the generated large WGS7DS and 21TRICH SNP matrices can be queried by subset at ICGRC CannSeek database (https://icgrc.info/genotype_viewer). The HASCH SNPs validation dataset is in Supple-mentary file 11, and the QTL analysis input is in Supplementary file 12, The Python script and sample inputs to generate the LP file from vcf is available at (https://doi.org/10.5281/zenodo.11149359). The LP file is accepted by Gurobi Optimizer or SCIP Optimization Suite (www.scipopt.org). SCIP is open source, while the user should obtain their own license to download Gurobi which we observed to run faster.All variant data for this study have been further deposited in the European Variation Archive (EVA) at EMBL-EBI under accession number PRJEB78836 (https://www.ebi.ac.uk/eva/?eva-study=PRJEB78836).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Southern Cross Plant Science, Faculty of Science and Engineering, Southern Cross University, 1 Military Road, East Lismore, NSW 2480, Australia
[2]International Rice Research Institute, Pili Drive, Los Banos, Laguna, Philippines

Received: 20 February 2024 / Accepted: 22 August 2024
Published online: 29 August 2024

References

1. Fordjour E, Manful CF, Sey AA, Javed R, Pham TH, Thomas R, et al. Cannabis: a multifaceted plant with endless potentials. Front Pharmacol. 2023;14:1–36.
2. Calvi L, Pentimalli D, Panseri S, Giupponi L, Gelmini F, Beretta G, et al. Comprehensive quality evaluation of medical Cannabis sativa L. inflorescence and macerated oils based on HS-SPME coupled to GC–MS and LC-HRMS (q-exactive orbitrap®) approach. J Pharm Biomed Anal. 2018;150:208–19.
3. Farinon B, Molinari R, Costantini L, Merendino N. The seed of industrial hemp (Cannabis sativa l.): nutritional quality and potential functionality for human health and nutrition. Nutrients. 2020;12:1–60.
4. Da Porto C, Decorti D, Tubaro F. Fatty acid composition and oxidation stability of hemp (Cannabis sativa L.) seed oil extracted by supercritical carbon dioxide. Ind Crops Prod. 2012;36:401–4.
5. Schluttenhofer C, Yuan L. Challenges towards revitalizing hemp: a multifaceted crop. Trends Plant Sci. 2017;22:917–29.
6. Thomson MJ. High-throughput SNP genotyping to accelerate crop improvement. Plant Breed Biotechnol. 2014;2:195–212.
7. Ramkumar G, Prahalada GD, Hechanova S, Lou, Vinarao R, Jena KK. Development and validation of SNP-based functional codominant markers for two major disease resistance genes in rice (O. sativa L). Mol Breed. 2015;35:1–11.
8. Kurokawa Y, Noda T, Yamagata Y, Angeles-Shim R, Sunohara H, Uehara K, et al. Construction of a versatile SNP array for pyramiding useful genes of rice. Plant Sci. 2016;242:131–9.
9. Begum H, Spindel JE, Lalusin A, Borromeo T, Gregorio G, Hernandez J, et al. Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). PLoS ONE. 2015;10:1–19.
10. Chen H, Xie W, He H, Yu H, Chen W, Li J, et al. A high-density snp genotyping array for rice biology and molecular breeding. Mol Plant. 2014;7:541–53.

11. McCouch SR, Wright MH, Tung CW, Maron LG, McNally KL, Fitzgerald M et al. Open access resources for genome-wide association mapping in rice. Nat Commun. 2016;7.
12. Thomson MJ, Singh N, Dwiyanti MS, Wang DR, Wright MH, Perez FA et al. Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications. Rice. 2017;10.
13. Kretzschmar T, Mbanjo EGN, Magalit GA, Dwiyanti MS, Habib MA, Diaz MG, et al. DNA fingerprinting at farm level maps rice biodiversity across Bangladesh and reveals regional varietal preferences. Sci Rep. 2018;8:1–13.
14. Sato M, Hosoya S, Yoshikawa S, Ohki S, Kobayashi Y, Itou T, et al. A highly flexible and repeatable genotyping method for aquaculture studies based on target amplicon sequencing using next-generation sequencing technology. Sci Rep. 2019;9:1–9.
15. Saikia M, Burnham P, Keshavjee SH, Wang MFZ, Heyang M, Moral-Lopez P, et al. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. Nat Methods. 2019;16:59–62.
16. van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR et al. The draft genome and transcriptome of Cannabis sativa. Genome Biol. 2011;12.
17. Laverty KU, Stout JM, Sullivan MJ, Shah H, Gill N, Holbrook L, et al. A physical and genetic map of Cannabis sativa identifies extensive rearrangements at the THC/CBD acid synthase loci. Genome Res. 2019;29:146–56.
18. Grassa CJ, Weiblen GD, Wenger JP, Dabney C, Poplawski SG, Motley ST, et al. A new Cannabis genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. New Phytol. 2021. https://doi.org/10.1111/nph.17243.
19. Lynch RC, Vergara D, Tittes S, White K, Schwartz CJ, Gibbs MJ, et al. Genomic and Chemical Diversity in Cannabis. CRC Crit Rev Plant Sci. 2016;35:349–63.
20. McKernan KJ, Helbert Y, Kane LT, Ebling H, Zhang L, Liu B, et al. Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes. bioRxiv. 2020. https://doi.org/10.1101/2020.01.03.894428.
21. Ren G, Zhang X, Li Y, Ridout K, Serrano-Serrano ML, Yang Y et al. Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa. Sci Adv. 2021;7.
22. Woods P, Price N, Matthews P, McKay JK, August. https://doi.org/10.1093/g3journal/jkac209
23. Medicinal Genomics. Kannapedia. 2024. https://www.kannapedia.net. Accessed 1 Dec 2023.
24. Zager JJ, Lange I, Srividya N, Smith A, Markus Lange B. Gene networks underlying cannabinoid and terpenoid accumulation in cannabis. Plant Physiol. 2019;180:1877–97.
25. Booth JK, Yuen MMS, Jancsik S, Madilao LL, Page AJE. Terpene synthases and terpene variation in cannabis sativa1[OPEN]. Plant Physiol. 2020;184:130–47.
26. Livingston SJ, Quilichini TD, Booth JK, Wong DCJ, Rensing KH, Laflamme-Yonkman J, et al. Cannabis glandular trichomes alter morphology and metabolite content during flower maturation. Plant J. 2020;101:37–56.
27. Braich S, Baillie RC, Jewell LS, Spangenberg GC, Cogan NOI. Generation of a Comprehensive Transcriptome Atlas and Transcriptome dynamics in Medicinal Cannabis. Sci Rep. 2019;9. https://doi.org/10.1038/s41598-019-53023-6.
28. McPartland JM, Hegman W, Long T. Cannabis in Asia: its center of origin and early cultivation, based on a synthesis of subfossil pollen and archaeobotanical studies. Veg Hist Archaeobot. 2019;28:691–702.
29. Woods P, Campbell BJ, Nicodemus TJ, Cahoon EB, Mullen JL, McKay JK. Quantitative trait loci controlling agronomic and biochemical traits in Cannabis sativa. Genetics. 2021;219.
30. Asamizu E, Ichihara H, Nakaya A, Nakamura Y, Hirakawa H, Ishii T, et al. Plant genome database Japan (PGDBj): a portal website for the integration of plant genome-related databases. Plant Cell Physiol. 2014;55:1–7.
31. Weiblen GD, Wenger JP, Craft KJ, ElSohly MA, Mehmedic Z, Treiber EL, et al. Gene duplication and divergence affecting drug content in Cannabis sativa. New Phytol. 2015;208:1241–50.
32. Xu Y, Li P, Yang Z, Xu C. Genetic mapping of quantitative trait loci in crops. Crop J. 2017;5:175–84.
33. Mandolino G, Carboni A, Forapani S, Faeti V, Ranalli P. Identification of DNA markers linked to the male sex in dioecious hemp (Cannabis sativa L). Theor Appl Genet. 1999;98:86–92.
34. Lubell JD, Brand MH. Foliar sprays of silver thiosulfate produce male flowers on female hemp plants. Horttechnology. 2018;28:743–7.
35. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 2011;6:1–10.
36. Murray KD, Borevitz JO. Axe: Rapid, competitive sequence read demultiplexing using a trie. Bioinformatics. 2018;34:3924–5.
37. GBS-PreProcess. 2018. https://github.com/relshire/GBS-PreProcess
38. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE. 2014;9.
39. Mansueto LGATK, Parabricks Gadi. Benchmarking. 2022. https://doi.org/10.5281/zenodo.8348884
40. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017;33:2037–9.
41. ICGRC. ICGRC Portal. 2022. https://icgrc.info
42. Gurobi. Gurobi Optimizer. 2008. https://www.gurobi.com/solutions/gurobi-optimizer
43. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
44. Arthur R, O'Connell J, Schulz-Trieglaff O, Cox AJ. Rapid genotype refinement for whole-genome sequencing data using multi-variate normal distributions. Bioinformatics. 2016;32:2306–12.
45. Broman KW, Wu H, Sen Ś, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19:889–90.
46. Taylor J, Butler D. R package ASMap: efficient genetic linkage map construction and diagnosis. J Stat Softw. 2017;79.
47. Lincoln SE, Lander ES. Systematic detection of errors in genetic linkage data. Genomics. 1992;14:604–10.
48. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A. 1987;84:2363–7.
49. Voorrips RE. Mapchart: Software for the graphical presentation of linkage maps and QTLs. J Hered. 2002;93:77–8.
50. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF et al. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet. 2013;9.
51. Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A. Feature selection methods and genomic big data: a systematic review. J Big Data. 2019;6.
52. Shah SC, Kusiak A. Data mining and genetic algorithm based gene/SNP selection. Artif Intell Med. 2004;31:183–96.
53. Grinberg NF, Orhobor OI, King RD. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. Mach Learn. 2020;109:251–77.
54. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet. 2005;37:1217–23.
55. Barrett JC, Fry B, Maller J, Daly MJ, Haploview. Analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.
56. Chen Y-H. An Integer Programming Approach for the selection of tag SNPs using multi-allelic LD. Commun Inf Syst. 2009.
57. Broad Institute. Tagger. 2005. https://software.broadinstitute.org/mpg/tagger/
58. Nishiyama S, Sato K, Tao R. Integer programming for selecting set of informative markers in paternity inference. BMC Bioinformatics. 2022;23:1–17.
59. Gusfield D. Integer Linear Programming in Computational and Systems Biology. Cambridge University Press; 2019.
60. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3-new capabilities and interfaces. Nucleic Acids Res. 2012;40:1–12.
61. Natsume S, Oikawa K, Nomura C, Ito K, Utsushi H, Shimizu M, et al. V-primer: software for the efficient design of genome-wide InDel and SNP markers from multi-sample variant call format (VCF) genotyping data. Breed Sci. 2023;73:415–20.
62. Yuan J, Yi J, Zhan M, Xie Q, Zhen TT, Zhou J, et al. The web-based multiplex PCR primer design software Ultiplex and the associated experimental workflow: up to 100- plex multiplicity. BMC Genomics. 2021;22:1–17.
63. O'Halloran DM, PrimerMapper. High throughput primer design and graphical assembly for PCR and SNP detection. Sci Rep. 2016;6:1–10.
64. Arbelaez JD, Dwiyanti MS, Tandayu E, Llantada K, Jarana A, Ignacio JC et al. 1k-RiCA (1K-Rice Custom Amplicon) a novel genotyping amplicon-based SNP assay for genetics and breeding applications in rice. Rice. 2019;12.
65. Broman KW. Genetic map construction with R/qtl. Univ Wisconsin-Madison Dep Biostat Med Inf Tech Rep. 2010;214:1–41.
66. Shirasawa K, Oyama M, Hirakawa H, Sato S, Tabata S, Fujioka T, et al. An EST-SSR linkage map of raphanus sativus and comparative genomics of the brassicaceae. DNA Res. 2011;18:221–32.

67. Yu G, Cui Y, Jiao Y, Zhou K, Wang X, Yang W, et al. Comparison of sequencing-based and array-based genotyping platforms for genomic prediction of maize hybrid performance. Crop J. 2023;11:490–8.
68. Guo Z, Wang H, Tao J, Ren Y, Xu C, Wu K et al. Development of multiple SNP marker panels affordable to breeders through genotyping by target sequencing (GBTS) in maize. Mol Breed. 2019;39.
69. Semagn K, Babu R, Hearne S, Olsen M. Single nucleotide polymorphism genotyping using Kompetitive Allele specific PCR (KASP): overview of the technology and its application in crop improvement. Mol Breed. 2014;33:1–14.
70. Takeshima R, Ogiso-Tanaka E, Yasui Y, Matsui K. Targeted amplicon sequencing + next-generation sequencing–based bulked segregant analysis identified genetic loci associated with preharvest sprouting tolerance in common buckwheat (Fagopyrum esculentum). BMC Plant Biol. 2021;21:1–13.
71. Nagano S, Hirao T, Takashima Y, Matsushita M, Mishima K, Takahashi M et al. SNP genotyping with target amplicon sequencing using a multiplexed primer panel and its application to genomic prediction in Japanese Cedar, Cryptomeria japonica (L.f.) D.Don. Forests. 2020;11.

Publisher's note