Building a community-driven bioinformatics platform to facilitate *Cannabis sativa* multi-omics research

Authors:

Locedie Mansueto [1], Tobias Kretzschmar[1], Ramil Mauleon[1,2], Graham J. King[1]

[1]Southern Cross University, Military Road, Lismore New South Wales 2480, Australia

[2]International Rice Research Institute, Pili Drive, Los Baños Laguna 4031, Philippines.

Abstract

Global changes in Cannabis legislation after decades of stringent regulation, and heightened demand for its industrial and medicinal applications have spurred recent genetic and genomics research. An international research community emerged and identified the need for a web portal to host Cannabis-specific datasets that seamlessly integrates multiple data sources and serves omics-type analyses, fostering information sharing.

The Tripal platform was used to host public genome assemblies, gene annotations, QTL and genetic maps, gene and protein expression, metabolic profile and their sample attributes. SNPs were called using public resequencing datasets on three genomes. Additional applications, such as SNP-Seek and MapManJS, were embedded into Tripal. A multi-omics data integration web-service API, developed on top of existing Tripal modules, returns generic tables of sample, property, and values. Use-cases demonstrate the API's utility for various -omics analyses, enabling researchers to perform multi-omics analyses efficiently.

Availability and Implementation: The web portal can be accessed at www.icgrc.info

Subjects: Software and Workflows, Bioinformatics, Agriculture

STATEMENT OF NEED

Background on Cannabis

Cannabis sativa (cannabis) is one of the earliest domesticated crops for food, fiber and medicine [1]. As hemp, Cannabis was a significant source of fiber in ropes traditionally used for shelter, clothing, traps, harnesses, sails and halyards. Its seeds (hempseed) are relatively rich in nutritious oils of high omega-3 and omega-6 fatty acid content [2], proteins, and contain a variety of micronutrients for human diet [3]. Its female flowers produce a mixture of secondary metabolites, some of which have psychotropic effects and have been used in religious rituals and recreational use. Because of these properties and uses, cannabis was also regarded as an ancient 'wonder plant' coined with various

polarizing names including 'devil's weed', 'penicillin of Ayurvedic medicine', 'donator of joy' [4]. The secondary metabolites found in high abundance in glandular trichomes on the surface of female flowers for pathogen resistance [5], insect deterrent [6], and UV protection [7] also have pharmacological and nutritional potential [8]. Cannabis glandular trichomes produce large amount of terpenoids, and cannabinoids. Calvi et al. [9] identified 554 compounds in cannabis including 113 cannabinoids and 120 terpenes. More than 200 metabolites from different genotypes are volatile, with 58 monoterpenes and 38 sesquiterpenes being reported. The most abundant monoterpenes are limonene, b-myrcene, a-pinene and linalool, while sesquiterpenes are E-caryophyllene, b-farnesene, and b-caryophyllene.

Cannabis is dioecious and diploid with 9 autosomes and a pair of sex chromosomes (female XX, male XY). The diploid genome size estimated by flow cytometry is 1636Mb for female (818Mb haploid size) and 1683Mb (haploid size 843Mb) for male plants [10]. The first draft sequence reported a haploid size of 534 Mb [11], while the accepted standard reference genome to date is 876Mb [12] Cannabis had been classified as narcotic for most of the 20th century (Mead, 2014), but recent revisions of legal status and recognition of defined medicinal benefits has supported it as an emerging economic crop with the World Health Organization scheduling recommendation [13]. Improving cannabis for the food, fiber and medicinal traits has thus recently been the focus of both public and private sector research (Small, 2017). The global market projection for medicinal cannabis is $176 billion by 2030 [14]. At 21% cumulative growth rate, the global cannabis cultivation market including hemp, could reach USD 1.8T by 2030 [15].

The need for Cannabis multi-omics database and analyses platform

The assembly of reference genomes [12,16–19] for hemp and medicinal types has facilitated an abundance of publications and associated publicly available genomic, transcriptomic, proteomic, and metabolomic data sets. Data are available from general web portals also include cannabis genomic and other omics data (e.g. NCBI, ENA, DDBJ, CoGe)[20–22], plant-specific databases (e.g. EnsemblPlant, Phytozome, PlantCyc)[23–26], and Cannabis-specific databases. Kannapedia [27] is a commercial site from Medicinal Genomics, which provides summaries on the genomic variation of cannabinoid biosynthesis genes for nearly two thousand cultivars. Bulk downloads of most of the data, however, are not available as a built-in feature. CannabisDraftMap.org [28] provides an inventory of raw proteomics and metabolomics data. The site has no analysis tools or results available, and only features short descriptions of various cannabis genomics projects. The Cannabis Genome Browser [29] hosts the results from the first cannabis draft sequence publication [11] but has not been updated since. Leafly [30] is a commercial site with cannabinoid and qualitative description of terpenes. SeedFinder [31] provides an interface for cannabis breeders, seedbanks, growers, and enthusiasts by

collecting and standardizing information about cannabis. As of February 2024, the database has 31,823 different cultivars listed. While it is not an online shopping platform, it is supported by related advertisements. CannabisGDB [32] is an integrated cannabis functional genomics database with four modules: varieties, gene loci, metabolites, and proteins. 'Varieties' provides description for 8 assembled genomes available from NCBI. 'Gene loci' contains their in-house gene predictions and functional annotation on these 8 genomes, and these are displayed using JBrowse [33]. 'Proteins' includes the UniProt plant proteins matched to the in-house protein model. 'Metabolite' displays various bar charts showing percent cannabinoid content, as collected from external publications.

Most of the Cannabis-specific websites and databases described above host proprietary data; hence the focus is mainly on specific cultivars and trait characterization. Some have omics-type data but are usually from unpublished or proprietary sources. These limitations provided the motivation for this project to build an authoritative 'open-science' reference bioinformatics portal to support cannabis multi-omics research with capabilities that we believe will be of great benefit to the research community, including: (a) tools to perform quick analyses using the loaded datasets, (b) tools allowing data/contents curation by the community (c) creation and promotion on the use of research data standards, and (d) a platform for collaboration and communication by the cannabis research community.

The features and capabilities were initially intended to fulfill the needs defined by the International *Cannabis* Genomics Research Consortium (ICGRC). Formed during the 2020 Plant and Animal Genome conference by an international community of cannabis researchers from the public and private/commercial sectors, ICGRC aims to provide an authoritative source of rigorous and peer-reviewed scientific information relating to Cannabis genetics, genomics and phenotyping. The scope of ICGRC activities encompass (a) Information and data sharing, (b) establishing standards and reference information, (c) facilitating the unambiguous description and availability of genetic resources, (d) contributions to consistent annotation of gene identity and function, and (e) describing common genetic resources and where necessary, virtual gene-banks [34].

IMPLEMENTATION

Tripal chosen as technology for the Cannabis omics Platform

There is a range of widely-adopted systems in the public domain suitable for genomic and genetic web databases/portals, such as InterMine [35], MetaCyc [36], EnsemblPlant [24], and Tripal [37]. Tripal is a content management system for genomics data based on the Drupal platform (drupal.org). The data architecture is based on Chado [38], a highly normalized schema that uses biological ontologies to define bioentities and their relationships. Tripal architecture is modular, allowing plug-in features such

as ready-to-use specialized 'omics modules for loading, visualization, and analysis, built-in support for user-driven content management, and is GMOD [39] standards-compliant. Tripal is used in many plant genomic research communities [37,40], and various genome hub projects [41] including several from the SouthGreen bioinformatics platform [42]. Its major advantage is flexibility for community-wide contribution and curation of content, thus Tripal was chosen as the software platform for the project. Tripal setup for the project is described in protocols.io, referred to as protocol from onwards (reserved DOI: dx.doi.org/10.17504/protocols.io.n2bvj3nz5lk5/v3)[43], of which the major steps are: Cannabis dataset download, reanalysis, and formatting for Tripal loading (protocol steps 1-8), Tripal docker container instantiation (protocol step 9) and loading of the reference genomes and annotation (protocol step 10). Modules have then been installed, enabled and loaded with the corresponding data: gene_search (protocol step 11), expression (protocol step 12), phenotype (protocol step 13), map (protocol step 14), and finally synteny (protocol step 15).

Cannabis –omics, genetic, and phenotypic datasets in the public domain

Published -omics datasets are sourced from databases such as NCBI Sequence Read Archive (SRA) (RRID:SCR_004891) (raw sequences)[44], NCBI RefSeq (RRID:SCR_003496) (genome assemblies and annotation)[45], Gene Expression Omnibus (GEO) (RRID:SCR_005012) (gene expression omnibus)[46], UniProt (RRID:SCR_002380) (protein sequences and annotations)[47], OBO Foundry biological ontologies and controlled vocabulary OBO (RRID:SCR_007083) [48], Crop Ontology Crop Ontology (RRID:SCR_010299) [49] crop-specific controlled vocabulary,. Datasets have also been manually curated from the published article tables, figures, and supplementary data and sites, which are common sources, especially for non-model or less studied species like cannabis.

Genome assemblies and annotations

The first peer-reviewed publication for cannabis genomes were for Purple Kush (PK), a medicinal type (THC-dominant) marijuana cultivar, and Finola (FN), a hemp-type cultivar developed for oil seed [11]. Both PK and Finola assemblies were updated and improved in Laverty et al. [16] . CBDRx (cs10) is a CBD-type cultivar derived from a cross from Skunk#1 (marijuana) with Carmen (hemp) and is estimated to be 89% marijuana and 11% hemp by admixture analysis. This was the first Cannabis assembly with NCBI RefSeq gene annotations (cs10, GCF_900626175.2) for three years and used by many publications for gene assignment. Pink Pepper (ASM2916894v1) is the second assembly annotated by RefSeq just recently (November 2023).

Three primary genome assemblies are currently loaded to the portal: cs10 (aka CBDRx, a CBD-dominant line) [12], PK and FN [16]. They were chosen to represent the major cannabis end-uses and have transcript sequences for gene annotation. Other Cannabis genome assemblies currently in the

public domain, such as cultivars First Light [12], Jamaican Lions [19], Cannbio-2 [18], JL wild [17], and Pink Pepper were not included in this release of the portal.

The annotations from RefSeq for cs10, and our FINDER [50] annotation for PK and FN, were loaded and searchable using the 'Gene Function Search' module interface and displayed in the JBrowse browser (RRID:SCR_001004). The query constraints are cultivar/assembly, annotation analysis, genome region, gene or transcription accessions, and keyword in the functional annotations. Query by sequence is also possible using BLAST (RRID:SCR_004870) interface. A summary of annotations available in the portal is in Table 1.

Transcriptomes, comparative expression, and proteomics datasets

The public availability of transcript data has been the primary consideration for our choice of genomes to load in the portal (Table 2). For comparative analyses either absolute expression levels across multiple samples or relative fold-change of transcript levels between two samples is available. Currently three (3) Cannabis transcript assemblies are publicly available: canSat3, finola1 and cannbio2. All were used to annotate their respective genome, and in differential expression studies across tissues [11,18,51]. These transcript assemblies are viewable in the JBrowse browser. Functional annotation was assigned for each transcript according to homology with RefSeq mRNA sequences, and the genome location by splice-aware alignment with the different genomes. As with gene loci, the transcript assemblies can be queried in the 'Gene Function Search' page, and their locations and transcript levels are also available to be displayed in JBrowse.

Comparative expression studies for Cannabis are limited, mostly on changes related to secondary metabolism and fiber synthesis, comparing different cultivars, treatment conditions, tissues or developmental stages. A gene expression study for bast fiber development in the textile hemp variety Santhica 27 included tissues sampled from the bottom, mid and top stem region [57] carried out. In another study, Finola trichomes were separated from mid-stage flowers and categorized into bulbous, sessile, and mature stalks [52]. Terpene and cannabinoid profiles were measured and correlation with synthase gene transcript levels assessed [53,58]. Relevant expression datasets in NCBI GEO were also loaded. The list of absolute and differential expression datasets available within the ICGRC Tripal resource is shown in Table 3 and Table 4 respectively. Expression data are displayed as transcript heatmaps of genes by samples. It is important for this module to have traceable assignment of sample attributes from the source. Expression data are also displayed overlaid on the generic Biopathways module.

As of completion of this project, only one Cannabis proteome expression dataset has been available in GEO. Using quantitative time of flight mass spectrometry, 1240 proteins from glandular

trichomes heads, 396 from trichome stalks, and 1682 from whole flower tissue isolates were identified by Conneely et al. [61].

Variants discovery using resequencing datasets

Details on the reanalysis of Cannabis Next Generation Sequencing NGS datasets are described in an accompanying paper (DOI: 10.25918/preprint.367) about the CannSeek SNPs database [62]. Data available in NCBI SRA as of December 2022 consisting of 26 trichome RNA-Seq, 383 genomic DNA, and the Phylos amplicons were reanalyzed using the GATK (RRID:SCR_001876) [63] and Parabricks [64] variant calling pipelines. The results are displayed using the CannSeek software, derived from Rice SNP-Seek [65], hosted in a Podman, or Docker, container [66,67] and embedded as a Tripal page.

Maps

The interactive *Map Viewer* , a part of the Tripal platform [68], can display the entire genome or chromosome, and features such as markers and QTLs in a single view. Maps positions may either be physical (use base position - bp), or genetic (use centi-Morgan - cM). Common markers are linked between the genetic and physical maps when displayed side by side. Public Cannabis genetic linkage data are scarce; only two recent publications of QTL maps are currently loaded; a study on gene duplication and divergence affecting drug content (Weiblen et al 2015) and QTL mapping for agronomic and biochemical traits (Woods et al. 2021).

Phenotypes

The *Phenotype Viewer* displays the distribution of measured traits including metabolites (cannabinoid and terpenoid content and their relative composition) across samples and experiments, visualized in a 'violin plot' [69] . Table 5 lists the phenotypic datasets loaded. This feature requires three genus-specific controlled vocabularies for trait, method, and unit. During setup, they can be defined from existing ontologies or community customized. New terms may be added as new datasets are loaded. Currently these ontologies use ChEBI [73] for chemical traits, PECO [74] for method and UO [75] for units. In the absence of a cannabis-specific vocabulary for methods and units in Crop Ontology, publication-specific terms were created when no suitable match in the existing ontologies were found.

Preparing Datasets for the Tripal modules

Gene model prediction and annotations

All resequencing data were trimmed using Trimmomatic (RRID:SCR_011848) [76] (protocol step 1). Cs10 genes and annotations were downloaded from NCBI RefSeq. (protocol step 2) The FINDER [50] pipeline was used to predict gene models for updated assemblies of Purple Kush and Finola using cultivar-specific RNA-Seq data. GffRead (RRID:SCR_018965) [77] was used to cluster overlapping gene

loci predictions into a common locus. (protocol step 3) Proteins were annotated for putative function/structure/location by homology search against the UniProt [47] databases using mmseqs2 [78]. Gene Ontology (RRID:SCR_002811) and InterPro (RRID:SCR_006695) IPR protein domains annotation of these protein were done using InterProScan (RRID:SCR_005829) [79]

Gene expression

Expression data from GEO [46] were loaded using the loader for expression and biosample. To build a system that hosts and integrates multiple -omics datasets for the same sample sets, secondary analyses were performed on public data to complete these for all -omics data types. Metabolite measurements were loaded and RNA-Seq sequences were reanalyzed. Gene expression using cs10 mRNAs were quantified on 21 trichome-specific public RNA-Seq datasets [51–54] using Salmon (RRID:SCR_017036) (protocol step 5). Transcripts from the 21-trichomes RNA-Seq samples were assembled following (protocol step 4) using rnaSPAdes (RRID:SCR_016992) [80], then mapped to the three references using GMAP (RRID:SCR_008992) [81] and ORFs were identified using TransDecoder (RRID:SCR_017647) [82]. Quantification of transcript abundance was then performed using Salmon following (protocol step 5).

Synteny block discovery

This follows the Tripal synteny viewer instructions of generating synteny blocks using gene annotations (protocol step 8). BLASTN (RRID:SCR_001598) and MCScanX (RRID:SCR_022067) [83] were used on the cs10 gene annotation and our generated predictions for Purple Kush and Finola.

BLAST database

BLAST databases for the 3 genomes and the annotated genes are installed in the portal, allowing genes and protein sequences from the users to be searched against genome assembly, coding sequence and proteins for cs10, Purple Kush PK, and Finola FN. Cs10 gene model sequences are from NCBI Refseq, while PK and FN are from the FINDER gene prediction pipelines.

Data preparation for external, non-Tripal tools

Some data are hosted and visualized using modules that are not part of Tripal (e.g. genome browser, pathways visualization, genotype management). Although Tripal might have some equivalent module, or the module might be implemented differently in other Tripal sites, the following tools were used for user benefit.

JBrowse Genome Browser

The JBrowse Genome Browser [33] displays genome features along the reference chromosome or scaffolds. The data is loaded from interval files like GFF or BED. It is highly interactive and web-based making it popular and used by many genomic database sites. Genes and transcript alignments for cs10, PK and FN genomes are available.

Biopathways mapping

MapManJS, a web version of MapMan (RRID:SCR_003543) [84], is a gene expression and biological pathway visualization tool. It uses curated, organism-specific sets of genes for each pathway. (protocol step 7) Since MapMan does not have Cannabis-specific pathways, we lifted over (via best match with Cannabis cs10 genes using mmseq2 reciprocal best hit on proteins) genes and pathways from other plant maps, specifically *A. thaliana*, tomato and eucalyptus. The expression of the cs10 genes is visualized over the selected pathway mapping and study loaded in the Expression module.

SNPs discovery and genotype data management

Variant data of the 21 trichomes RNA seq were discovered using the GATK-RNASeq pipeline (protocol step 6). Genomic resequencing data for 398 samples were also used to discover SNP variants using the GATK germline SNP calling workflow [85] and Parabricks [64].

Genotype data results from variant discovery pipelines are presented as large genotype matrix of samples and base position. The Tripal Natural Diversity Genotypes (DOI:10.5281/zenodo.3731337) module used in KnowPulse [69] is able to display this matrix. However, it uses web technologies that limit its interactivity and has performance issues on larger genotype matrix queries. The Rice SNP-Seek database [65] can display thousands of SNP positions and samples while maintaining interactivity and allows more analysis on the result and is used in this project.

RESULTS

Setup and customizations of the Tripal platform, and re-engineering modules for multi-omics data integration

The detailed workflow to set up and customize the platform is enumerated in the protocol [86], and illustrated in *Figure 1*. Documentations about instantiating various Tripal modules and loading data are notoriously scattered across publications, and this workflow aims to collate and summarize the steps required to format files to simplify preparation of the various modules. The workflow generates the files necessary to initially populate and test the different Tripal modules. This can also guide installations of future Tripal sites, especially for plants with scarce published data, or where only raw sequences are available.

We added features that are not part of Tripal that were essential to enable multi-omics analysis, including pathway and expression visualization (MapMan JS), and genotype mining and visualization (SNP-Seek). NCBI BLAST was also included in the platform. (protocol step 18) provides details on the installation of these external applications (JBrowse, CannSeek, MapManJS, BLAST).

In addition, we developed a novel module specifically for multi-omics integration. This is in the form of a web-service with a predefined Application Programming Interface (API) that queries and merges

the different Tripal modules, facilitating rapid retrieval of different data types from multiple sources. The capabilities of this API are demonstrated with use-cases typically performed in -omics studies.

Challenges encountered in the existing Tripal and Chado design and solutions implemented

Although existing Tripal modules can host and display various biological -omics datasets, integrating data from the different modules is not straightforward. The underlying Chado schema is highly flexible and normalized. Unlike the more common Entity-Relation, Chado follows the Entity-Attribute-Value Model database model[87]. The data model relies on biological ontologies, which are loaded as data, to define datatypes and relationships. The different loaders for each module populate database tables, which may be scattered, duplicated and confusing in their utility and terminologies.

An additional complication is that Tripal introduces a layer of abstraction for biological entities on top of the storage tables in the Chado schema. Using ontology terms as type identifiers, a Chado table can store records for different biological entity types. Table 6 lists the default Tripal entities and their corresponding Chado tables and types. The Tripal BioEntity objects and Chado relational table mappings are illustrated in Figure 2. The feature, analysis and stock tables are used in common by multiple Tripal BioEntity types. Data management can be challenging when mapping the relationships between distinct bioentities because they are not equivalent to the entity-relationship defined by the Chado schema. Another challenge is the requirement to manually synchronize the Chado records and the Tripal entities when loaders are not used There are synchronization buttons in Tripal to publish Chado records to Tripal, unpublish deleted records, or delete unpublished bioentities. Using these features requires thorough understanding and experience on the datasets and platform.

Several solutions are implemented to overcome the challenges encountered. The integration process is built on top of existing Tripal modules, specifically for the ‘chado_gene_search’, ‘expression’, and ‘phenotype’ modules. These modules generate materialized views from complex queries used in the visualizations. The -omics integration API queries these materialized views by SQL. Merging from the different modules is implemented using table pivot. All queries return a triplet of property, biomaterial and value. A pivot function then transforms them into a table with the biomaterials as column labels, and properties as row labels. Figure 3 illustrates how the modules are linked for integration. Biomaterial labels can be from ‘dbxref’, ‘biomaterial’ or ‘stock’ tables. Properties are from ‘cvterm’ (sample metadata, traits) or ‘feature’ (genes, transcripts, proteins). The values are collected from any of the other tables following their references. Controlled terms are defined in the ‘cvterm’ table, but their assignment to a specific biomaterial and property are in their datatype-specific tables. The generic -omics query result table should be able to handle any of the dataset types, factors, and variable and values in Table 7.

The datasets are linked by sample IDs defined by their NCBI BioSample whenever available. Publications are required to define SAMN sample IDs when they submit sequences to NCBI, including NGS genotyping and RNA-Seq studies.

We also identified several ambiguous usage of terminology while integrating the Tripal modules. Table 8 summarizes the exact definition of tables mentioned as defined in Chado, or BioEntities as defined in Tripal.

For the 'stock' and 'biomaterial' tables, these appear to hold similar data. However, from the Chado definition, 'stock' represents the physical entity of an organism identified by unique name and stock type that is held within a living or preserved collection. In contrast, 'biomaterial' is a biological material (tissue, cells or serum) that may have been processed. Biomaterials should be traceable to source raw biomaterial, recognizing that unless vegetatively cloned, individual plants of a dioecious outcrossing species are likely to have different genetic makeup (haplotype). The Expression module uses the Biomaterial type, while the Phenotype module uses the Germplasm Accession type to represent the samples. Merging expression with phenotype data needs a link between stock and biomaterial which is not explicitly defined in Chado. The workaround implemented in the project is to use the same property and value (NCBI BioSample ID) in the stockprop and biomaterialprop tables. However, BioSample (SAMN) IDs may not be available since only sequences submitted to NCBI or other INDSC repositories have them. In this situation we assigned stocks/biomaterial ID using the simplest unique name deducible from the source publication.

Issues in the 'organism' and 'genome assembly' entities were also encountered. Although the portal handles only *Cannabis sativa* L. as the target organism, the 'organism' entity is used in Chado and Tripal to define a set of contigs and genes for a given genome. Thus, the different genome assemblies for cs10, PK and FN are stored as distinct records in the organism table. In Tripal, the Genome Assembly entity is defined as a computational analysis for a genome assembly in the analysis table. As implemented in the project, we loaded the different assemblies as separate organisms in the organism table, then modified all Organism labels to Assembly.

For the entities 'Project', 'Study', 'Experiment' and 'Analysis', by Chado definition, "Study" represents an experiment that produced microarray data; "Project" is a flexible property table for projects; "Analysis" is a single unit of computational analysis; "Assay" consists of a physical instance of an array, when combined with conditions create an Array.

The different types of computational analysis are reflected in Tripal by using the specific analysis entities. Currently defined are Genome Annotation, Genome Assembly, BLAST Results, InterPro Results. The phenotype module loads the experiment (Experiment is used in the label) to the Project table and is associated to a genus and the samples to Germplasm Accessions. A genus is defined in the phenotype configuration to consist of a controlled vocabulary of Trait, Method and Unit, or optionally

by Crop Ontology. "Germplasm Accessions" is a type of stock in Tripal so phenotype samples are loaded to the stock table. Phenotype studies are added to the Project table. The Expression and Biomaterials loader in the expression module associates the data to an Analysis and Organism, which was used earlier for assemblies.

The interchanging use of terms made it confusing to merge samples between different modules, and our workaround was to use common terms for both the stocksprops and biomaterialprops table.

The Tripal genotype module, which uses the Natural Diversity tables [88], was tested and found to be performing sub optimally especially when loading large genotyping data matrices generated from NGS variant calling. To solve the speed issue, the SNP-Seek software system [65], developed to host large SNPs datasets, was instantiated in a podman container and embedded as a Tripal page. This feature, now named CannSeek database, is described in the accompanying paper. For the API calls to query SNPs, BCFtools (RRID:SCR_005227) [89] was used directly on the VCF files stored separately for each reference and dataset. The genotyped samples metadata are loaded to the 'biomaterial' table so they can be autoloaded from NCBI BioSample.

Some issues for the entities 'gene', 'mRNA', and 'protein' were also resolved. The central dogma of molecular biology relations is well-defined in the Tripal model and all are stored in the Chado feature table by different feature types. Their relationships are explicitly related as: mRNA part_of Gene, Exon part_of mRNA, Polypeptide derives_from mRNA. BLAST and InterPro Analysis Results were assigned to the specific molecule type used (i.e. DNA, mRNA or polypeptide). In the 'chado_gene_search' module, keyword search was modified to include the annotations assigned to all genes, mRNA and proteins by tracing their relationships. However, it became challenging to cross-reference genes between external databases and applications. The accessions or IDs will be different for the same gene but different molecule types. In addition, a gene locus can be associated with multiple mRNAs and proteins. To avoid the complicated query, a materialized view was created to map the corresponding genes, mRNA and polypeptides.

Replicates are challenging to handle when samples are measured, analyzed, and reported on multiple data types. Replicates may be biological, (i.e., multiple studies using the same plant source/genebank accession, multiple individuals in a study) or technical (i.e., multiple samples from a tissue with the same treatments). In NCBI, a biomaterial in BioSample (SAMN) can have multiple sequence replicates in SRA (SRR). Replicated NGS reads can be variant-called separately or jointly for a given SAMN. The measurement of metabolites for a SAMN can also have technical replicates. Some phenotype datasets loaded from publications report measurements per replicate [53], while others the mean and standard deviation [54] . The Tripal phenotype module documentation suggests loading

the individual replicate data [88]. However, their proposed use-case is focused on geolocational and seasonal factors, which may be more applicable for a single project or institute, but not for a community portal that integrates from various data sources. These replicates create a Cartesian join between biomaterials and properties, that is, rows or columns will be duplicated except on one element. For this reason, we use average and standard deviation for technical replicates from published datasets. We pre-calculated these statistics before loading when only replicate values were available. It is also possible to have an array of values for a given biomaterial and property without a traceable replicate number. For publications that report in this format, the system can still store them and report in JSON format.

The -omics integration APIs: database query generation and web service implementation

The normalized structure of Chado makes database SQL queries complex and inefficient. To circumvent this, the Tripal modules prepare materialized views of the necessary data ready for user query and visualization. The same strategy was used to integrate the datasets from the different modules where they shared common BioEntities. Details of the data integration queries are described in (protocol step 16). In brief, five datatypes are queried separately, each returning four columns (i.e. datatype, property, sample, value). The datatypes are ID, PROP, PHEN, EXP and SNP. ID refers to sample identifications, PROP to sample attributes and metadata, PHEN are phenotype variables, EXP are transcripts or protein IDs, and SNPs are SNP IDs. The results are merged by a UNION operation, pivoted into a DATATYPE, PROPERTY by SAMPLE table, and returned to the client.

The API for omics data, implemented as a web service, usually returns JSON format text. They are returned for separate experiments, datasets, or omics type like BrAPI [90], or Tripal Web-Service [91]. Here we return an efficient and intuitive table format ready for further processing, which can be loaded easily into spreadsheets or any csv reader library. This API will work in complement with the Tripal Web Service module, which is also enabled in the ICGRC site [92]. The web-services module can supply the metadata of the biological entities, but not the values of the different -omics measurements. Specifically, the values this API queries are measurements in the materialized views taken from the Chado tables elementresult.signal (expression), phenotype.value (phenotype), stockprop.value (stock property values) , biomaterialprops.value (biomaterial property values). The allele values for variant queries are read directly from the VCF files due to their large size if loaded to the Postgres database. The samples and metadata of the VCF are in the ‘biomaterial’ and ‘biomaterialprop’ tables.

To avoid Cartesian join duplicacy for replicates, complex element object representation like JSON was used for list of values. Being able to present values with replicates, their statistics, methods

and units as a table element, the table still has the equivalent representation power of a JSON object typical in web-services APIs.

The Cannabis Tripal Web portal

The instantiated platform is available at https://icgrc.info [93]. Figure 4 shows the main features of the site. The functional menu includes Genomics, Phenome, Tools, and Infome. 'Genomics' includes datasets pertaining to sequences, while 'Phenome' includes phenotypes and gene expression data. 'Infome' allows for management of Tripal Bioentities metadata, manuals, and collaborative information. The 'Genetic Resources' section is reserved for data related to gene banks and other genetic resource collections. The 'Tools' section is for performing computations using the datasets, currently BLAST is available. Details on the use are provided in the User's Manual in the menu.

Drupal administrative modules allow site and user administration. Users are assigned roles of public, registered, data curator, site curator, or administrator. Registered users can view additional information including the metadata of all samples, biomaterials, analyses, or any bioentities made private to a group of users. They can also post comments on some pages and data. To register, only a name, affiliation and a business email address are required to be provided.

Most of the modules retain standard Tripal functionality, but some were revised to expand their capability. In 'Gene function search', keyword search now includes all annotation assigned to gene, mRNA, or protein. Using a Gene Ontology accession (GO:nnnnnn) includes all sequence features annotated with any of the transitive closure of the GO term. The expression module has an added feature to generate a file intended for the Biological Pathway viewer. The 21-trichomes transcript expression levels were also converted to bigwig format and visualized as JBrowse tracks at the coordinates identified with TransDecoder. Lastly, unique to this portal are the MapMan biological pathway expression viewer [84], and the CannSeek SNPs viewer. The gene search materialized view SQL definition and MapManJS HTML page revisions are in the protocols.io steps 11 and 18, respectively.

The protocol (DOI:10.17504/protocols.io.n2bvj3nz5lk5/v3) developed can also serve for future site update, and in setting-up new Tripal sites especially for crops where only sequence datasets are readily available.

The -omics API query

We designed and implemented an API to serve programmatic queries to the omics datasets. The available queries and parameters are described in the ICGRC Omics API documentation [94]. The general API query format is,

https://icgrc.info/api/{user}/{datatype_list}/{constraints_list}?parameters=..&..&...

where {user} could be an authenticated user with login and privileges, or just 'user' for public access. {datatype_list} is a comma separated set of datatypes to query with possible values of [gene, phenotype, expression, variant]. {constraints_list} is a comma separated list of genomic regions [chromosome:start-end] or keywords. Genomic region is used when Variant is queried. Keywords are queried on the gene functional annotations, as used by the 'Gene Function Search' page, to get genes for Expression or Variant queries. The available datasets and summaries are available at **https://icgrc.info/api/user/{datatype}/dataset**. The results of queries with multiple datatypes are combined by SAMPLE. For example, for samples with both expression and phenotype data, the sample column will have both genes and traits/metabolite rows.

DISCUSSION

Standardization for Cannabis vocabulary and nomenclatures

Standardization is mandatory to effectively use the platform, when hosting diverse datatypes and sources. Here we summarize terms that require community-wide standardization, their impact in storage and analyses, the current best practices from various research groups, possible immediate workarounds, and suggestions to prompt the establishment of accepted community standards.

Sequence and sequence features accession IDs

NCBI assigns unique sequence identifiers for all datasets it hosts and analyses. Similar ID assignments are implemented by other databases like Ensembl, UniProt, etc. External references for these identifiers are handled in the 'dbxref' table in Chado.

However, prior to publication, different research groups tend to generate their own IDs and nomenclatures. It would be highly beneficial if a standard in generating nomenclatures was adopted by the community. The standard alphanumeric coding system may consider the increments, versioning, feature types, relationships between feature types, and relationships with other relevant ontologies.

Stocks, cultivars, biomaterials, sample

This is a prevailing gene bank issue [95]; unfortunately there is no authoritative gene bank for Cannabis [96]. NCBI assigns sample SAMN IDs for submissions in BioSample. However, these IDs are more applicable for biomaterials, specifically laboratory materials. They can't trace the pedigrees of genetic stocks.

In absence of a physical gene bank, the creation of a virtual gene bank was suggested [34]. A possible model for this is the Genesys platform[97]. Under this system, physical gene banks worldwide publish their passport and characterization data, which are then merged by Genesys into a common

query interface. This or a similar entity would then have the authority to assign Stock or Germplasm accessions for Cannabis.

NCBI BioSample metadata submissions

Related to biomaterials, the NCBI BioSample SAMN metadata can be loaded automatically into Tripal. Thus, terms used in NCBI submissions are reflected in the Tripal 'Biomaterial' pages. The sample metadata are also in the ID and PROP rows returned to API queries. It is therefore beneficial to use standardized terminologies during submissions to public repositories like NCBI, to use the information efficiently and accurately in subsequent analyses.

Querying PubMed for publications relating to cannabis plant biology or agronomy is complicated by a wealth of papers pertaining to clinical or pharmacological studies. We therefore recommend using standard keywords or MESH term specific to the study of the plant itself, not its clinical/pharmacological use.

Traits, variables, methods, units for Cannabis

Detailed crop ontologies are available for well-studied crops, although these have their limitations [98]. These terms are often derived from legacy dictionaries used by gene banks in their passport transactions and characterization of collections. A standardized Cannabis crop ontology would benefit both the industry and scientific community to facilitate data curation and management. Moreover, using them during submission to public repositories simplifies their integration as mentioned above.

Multi-omics query using the API: use-cases

With the different omics data types loaded in the database, we are able to integrate these using API queries. We created Jupyter (RRID:SCR_018315) notebooks and Python modules to demonstrate the potential uses of the API for multi-omics analyses, all available in (protocol step 17). A static export of the notebook with outputs is in the ICGRC Omics Demo [99]. In the notebook, the python modules (module names are in `courier font`) perform specific tasks and use Pandas (RRID:SCR_018214) Dataframe for data handling and manipulation. Depending on the intended analysis, additional software or modules need to be installed in the client as described in the notebook.

Use case 1: Get phenotypes, normalize units, plot distributions and batch effects correction

https://icgrc.info/api/user/phenotype/list. This query returns a table of phenotype values from multiple datasets, in mixed units and format. When merging values from multiple sources, compatible variables need conversion of units. Units and standard deviation information for each value are included with flag with_stdunits=1. The values with different units can then be converted to a uniform unit (`convert_units_to`), for plotting and further processing. The main purpose of this feature is mainly for storage, integration, and retrieval. For the statistical validity of merging data from

different sources and studies, we included a minimal feature (`check_batcheffects`) to detect and correct for batch effects and impute missing values using the samples metadata. This module uses ski-learn SimpleImputer and PCA [100], ppca [101] and pyComBat (RRID:SCR_010974)[102]. The metadata and sample attributes can also be used as covariates in various analyses.

Use case 2: Co-expression analysis using WGCNA

https://icgrc.info/api/user/expression,phenotype/list. This query returns a table with phenotype and gene expression values for each sample in the selected datasets. These data are used in Weighted correlation network analysis, WGCNA (RRID:SCR_003302) [103] in the (`do_wgcna_prepare`) and (`do_wgcna_modulesgenes`) modules. These modules return the top gene-trait pairs, with gene significance and p-value.

Use case 3: Matrix eQTL

https://icgrc.info/api/user/expression,variant/{querygenes}. This query returns a table of gene expressions and SNP allele data for genes and regions that meet the querygenes annotations constraints. The module (`do_matrixeqtl`) uses Matrix eQTL [104] for the top SNP-gene pairs by p-value. If provided with SNP and genes position information, Matrix eQTL can also analyze for local (cis) and distant (trans) pairs separately. The gene information is also queried using the API https://icgrc.info/api/user/gene, constrained by position for the SNPs, or by accession for the expression genes.

Use case 4: SNP-trait association study

https://icgrc.info/api/user/variant,phenotype/{querygenes}. This query returns a table of traits, mostly metabolite contents for cannabis, and SNP alleles for genes and regions in querygenes. The module (`do_gwas`) uses PLINK (RRID:SCR_001757)[105] –assoc function to get the top SNP-phenotype associations by p-value. When performed on all SNPs for a genome, and with enough samples this is basically a GWAS. To-date, there are only few public Cannabis samples that have both phenotype and genotype data publicly available, hence the limitation on datasets loaded in the portal. The -omics tools presented above used the API to retrieve multiple datasets for various analyses. Each analysis returns the top associations by p-value between genes, genotype variants, biomolecules, traits, or other factors. From the multiple evidence collected, it is logical next step to overlap the relationships into networks (union) or Venn diagrams (intersection) for further interpretation. The merge module (`merge_matrixeqtl_wgcna_gwas`) takes the results from WGCNA, eQTL, and plink -assoc, intersect the genes and traits, and merge the pairs to identify SNP gene-traits associations supported by two evidence, as illustrated in Figure 5.

The example use-cases illustrate the flexibility of the API for multi-omics, multi-source analyses. We showed two ways of merging multiple datasets. First is by merging datasets as input to an analysis

tool. This requires shared samples, batch effect correction, covariates information, and most likely heavy imputations. Their important link is the shared genotype. In the second approach, the input for an analysis tool is from one dataset, but the results from different analyses are merged by shared biomolecules. Each approach has its own advantages and applications. Integration of sequence data is performed in the backend during variant calling, transcript assembly, or gene expression quantification.

CONCLUSION

We setup the Tripal platform to host available Cannabis genomics data available in the public domain. We also provided a protocol to generate various datasets from sequence data, to load, and test the different modules and visualizations. The protocol simplifies the process of regularly updating the site. One year since our last upload, a recent check on NCBI and PubMed returned new publications and datasets for potential reanalysis and loading to the site. This shows that Cannabis research is active and rapidly growing. The setup of this web-portal is timely, and ready to serve the research community. We also highlight the importance of using community standards especially when depositing datasets in public repositories. Standards minimize third-party curation and encourage reuse for meta-analysis. The multi-omics integration API and the example Jupyter notebooks demonstrate the convenience of doing analysis when the data are readily available in table format. Since they are built on top of Tripal modules, they can be readily adapted to any Tripal site for other research communities.

AVAILABILITY OF SOURCE CODE AND REQUIREMENTS

The source codes in this project can be grouped into:

a. The ICGRC web application which is an instance of Tripal v3, the various Tripal/Drupal modules installed, site customizations/settings saved in the database, and the custom -omics API web server module developed in this project.
b. the scripts and command lines part of the computational protocol for dataset generation from public datasets.
c. Jupyter notebooks and python client library for -omics analyses which call the omics API

b. and c. are embedded or attached in the sections of the protocol described below. a. are available from the authors upon reasonable request.

Project name: **ICGRC Portal Tripal Data Generation and Setup**

Project home page: https://www.protocols.io/private/34114B4120C511EF83720A58A9FEAC02

Operating system(s): Linux

Programming language: bash, PHP, Python

Other requirements: listed in the steps in the protocol

DOI: dx.doi.org/10.17504/protocols.io.n2bvj3nz5lk5/v3 (reserved)

ABBREVIATIONS

API, Application Programming Interface; cs10, CBDRx cultivar assembly; FN, Finola cultivar assembly; GEO, Gene Expression Omnibus; ICGRC, International Cannabis Genomics Research Consortium; NCBI, National Center for Biotechnology Information; PK, Purple kush cultivar assembly; QTL, Quantitative Trait Loci; RefSeq, Reference Sequence Database; SAMN, BioSample accession; SNP, Single Nucleotide Polymorphism; SRA, Sequence Read Archive; SRR, SRA accession

DECLARATIONS

Ethical approval

The authors declare that ethical approval was not required for this type of research.

Competing Interests

None

Author contributions

LM implemented and modified the architecture and contents of the software and is primary author of the manuscript. RM supervised the software development and contributed to the development of the manuscript. TK supervised the overall development of the manuscript as project lead. GJK contributed to the development of the manuscript and provided linkage to the International Cannabis Genomics Research Consortium. All authors have proofread the manuscript.

Acknowledgements

The authors acknowledge the provision of computing and data resources provided by the Australian BioCommons Leadership Share (ABLeS) program. This program is co-funded by Bioplatforms Australia (enabled by NCRIS), the National Computational Infrastructure and Pawsey Supercomputing Centre.

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). The ICGRC and CannSeek web servers are hosted and funded by SCU.

TABLES

Table 1. Gene annotations available for various published cannabis genomes

Cultivar	Annotation	Type	Genome Assembly	RNA-Seq/transcripts dataset	Method	Reference
Cannabis sativa (CBDRx, cs10)	CS10	gene, mRNA	GCF_900626175.2		NCBI RefSeq	RefSeq
Cannabis sativa (Purple Kush, pkv5)	PKGMv1	gene, mRNA	GCA_000230575.5		GeneMark EP gene prediction, best hit blastp with NCBI NR	[11]
Cannabis sativa (Purple Kush, pkv5)	PKFDv1	gene, mRNA	GCA_000230575.5	PRJNA73819 (SRR352195, SRR352196, SRR352198, SRR352200, SRR352201, SRR352202, SRR352203, SRR352205, SRR352208, SRR352210)	FINDER This study (protocol steps 2-3)	This study [11]
Cannabis sativa (Finola, fnv2)	FNFDv1	gene, mRNA	GCA_003417725.2	PRJNA73819 (SRR351932,SRR351933) PRJNA483805 (SRR7630400-SRR7630408)	FINDER This study (protocol steps 2-3)	This study [11,52]
(CBDRx, cs10)	cs10g21t	gene, mRNA	GCF_900626175.2	21 Trichomes	This study (protocol step 4)	This study [51–54]
Cannabis sativa (Purple Kush, pkv5)	Csativa	mRNA		Assembled transcripts from canSat3_transcriptome-representative.fa.gz		[11]
Cannabis sativa (Finola, fnv2)	Csativa	mRNA		Assembled transcripts from finola1_transcriptome-representative.fa.gz		[11]
Cannabis sativa L.	Csativa	mRNA	no specific assembly	C sativa mRNAs from NCBI Genes		

Table 2. Transcript assemblies available from published cannabis studies

Dataset	Method	Samples	Datasource	Reference
canSat3	Aligned with cs10 and pkv5 assemblies using minimap2 in splice alignment mode	Purple Kush	canSat3_transcriptome-representative.fa.gz	[11]
finola1		Finola	finola1_transcriptome-representative.fa.gz	[11]
Cannbio-2		Cannbio-2	GIFP01000001:GIFP01064413	[51]
5 transcripts	1. Clustered (mmseq2 cm2 80% identity) transcript assemblies combined from 5 data sources 2. Aligned to cs10 and pkv5 assemblies using minimap2 in splice alignment mode	Purple Kush	canSat3_transcriptome-full.fa.gz	[11]
		Finola	finola1_transcriptome-full.fa.gz	[11]
		Cannbio-2	GIFP01000001:GIFP01064413	[51]
		Terpene synthase genes from various *C. sativa* cultivars	AB057805.1, AB164375.1, AB164375.1, AB292682.1, CcTHCAs, CsCBDAs, KY014554.1, KY014555.1, KY014556.1, KY014557.1, KY014558.2, KY014559.1, KY014560.1, KY014561.1, KY014562.1, KY014563.1, KY014564.1, KY014565.1, MK131289.1, MK131290.1, MK614217.1, MK801762.1, MK801763.1, MK801764.1, MK801765.1, MK801766.1, MN967470.1, MN967471.1, MN967472.1, MN967473.1, MN967474.1, MN967475.1, MN967477.1, MN967479.1, MN967480.1, MN967481.1, MN967481.1, MN967482.2, MN967483.1, MN967484.1, MnCBDAs-like, MT295505.1, MT295506.1, TPS11, TPS11-like, TPS12, TPS12-like, TPS13-like1, TPS13-like2, TPS14, TPS17, TPS20, TPS23, TPS24, TPS30-like, TPS32, TPS36, TPS37, TPS38, TPS40, TPS41, TPS42, TPS43, TPS44, TPS46, TPS47, TPS48, TPS49, TPS4-like, TPS50, TPS51, TPS52, TPS53, TPS55, TPS58, TPS59, TPS60, TPS61, TPS62, TPS63, TPS64, TPS6-like, TPS7_Finola, TPS8-like, TPS9, TPS9-like1, TPS9-like2	[53,55,58]
		Cannabinoid synthase genes from Argvana Heart and Tachllta cultivars	GHVF01000001-GHVF01000012, GHVG01000001-GHVG01000012	[56]
21 trichomes samples	This study (protocol step 4) 1. Fastq raw reads from trichome RNA-Seq experiments were assembled using rnaspades 2. Transcript assemblies were aligned to cs10 and pkv5	Sour Diesel, Canna Tsu, Black Lime, Valley Fire, White Cookies, Mama Thai, Terple, Cherry Chem, Blackberry Kush	PRJNA498707	[53]
		Afghan Kush,Blue Cheese,CBD Skunk	PRJNA599437	[58]

	assemblies using minimap2 in splice alignment mode	Haze,Chocolope, Lemon Skunk		
		Finola bulbous, sesille,stalked	PRJNA483805	[52]
		Cannbio-2 trichome, 4 development stages	SRR10600922, SRR10600916, SRR10600908, SRR10600906	[51]

Table 3. Absolute Gene/Transcript Expression data from published cannabis studies

Samples	Tissue/Condition	Treatment	Measurements	Reference
21 Trichome samples	21 Trichome	none	Cs10 mRNA	This study (protocols.io step 5)
21 Trichome samples	21 Trichome	none	cs10g21t	This study (protocols.io step 4-5)
C. sativa cv. Santhica 27	Three stem regions: top, middle, bottom	none	FN, finola1	[57] GEO GSE94156
Purple Kush, Finola, USO-31 SRP006678 SRP008673	SRP006678 mature leaf, immature leaf, flower buds, primary stem, young leaf, mature flower, stem-petioles, entire root SRP008673 young leaf, roots, shoots, stems, pre-flowers, early-stage flowers, mid-stage flowers	none	PK, canSat3	[59] GEO GSE93201
Hemp cultivar Yunma 1	- leaves, roots, stem bast, and stem shoots were collected separately. - Separate ABA treatment and collected after 3, 6 hrs - Equal samples pooled	drought stressed plants (DS), control (CK), abscisic acid	PK, reads aligned to Purple Kush AGQN00000000 and transcript assembly with Cufflinks	[60] GEO GSE56964
Finola	Trichome: bulbous, sessile, stalked	none	FN, finola1	[52]

Table 4. Differential gene/transcript expression data from published cannabis studies

Samples	Tissue	Measurements	Reference
Cannbio-2	- stem, root-tip, root-mid, leaf from freshly planted cutting, vegetative plant to reproductive plant. - floral bud tissues and trichomes were isolated from reproductive plants at 35, 42, 49 and 56 days after induction of flowering in the female plants - male vegetative leaf and reproductive tissues (pollen sacs) from the male strain	Cannbio2 TSA GIFP00000000	[51] PRJNA560453
Hemp No.8	Flower, glandular trichome head and stalk	cs10 peptide, hemp on uniport	[61]

Table 5. Phenotype data from published cannabis studies

Reference	Cultivars	Tissue	Measurements
[70]	(33 x 3 reps) accessions sourced from the Ecofibre	leaf	Delta9-tetrahydrocannabinol, cannabidiol, cannabigerol
[71]	(30) Alien Blues,Blue Cherry Pie,Blue Dream,Bohdi Tree,Cold Creek Kush,Crystal Cookies,Double Royal Kush,FLO,Grandaddy Purple,Green Crack 2015,Green Crack 2017,Holy Power,Juanita,Lavendar Jones 2015,Lavendar Jones 2017,Lavendar,Love Lace,Oracle,Platinum Buffalo,Platinum Gorilla,Platinum Scout,Purple Fat Pie,Romulin,RX,Skywalker,Sour Willie,Thunderstruck,Twisted Velvet,Walker Kush,White Widow	flower, leaf	Delta9-tetrahydrocannabinol, cannabidiol, cannabichromene, cannabigerol, Delta9-tetrahydrocannabivarin
[53]	(9) Sour Diesel, Canna Tsu, Black Lime, Valley Fire, White Cookies, Mama Thai, Terple, Cherry Chem, Blackberry Kush	bud	(27) Delta9-tetrahydrocannabinolic acid, Delta9-tetrahydrocannabinol, cannabidiolic acid, cannabidiol, cannabigerol, cannabinol, cannabichromene, Total_cannabinoids, beta-myrcene, limonene, alpha-pinene, beta-pinene, 1,8-cineole, linalool, terpinolene, borneol, beta-ocimene, camphene, gamma-3-carene, camphor, plus-terpinene,

			Total_monoterpenes, beta-caryophyllene, alpha-humulene, nerolidol, Total_sesquiterpenes, Total_terpenoids
[58]	Lemon Skunk Jack Herer Blueberry Chocolope Afghan Kush CBD Skunk Haze Vanilla Kush Blue Cheese	flower	(47) α-pinene, myrcene, green-leaf-volatile, limonene, (E)-β-ocimene, linalool, terpinolene, Monoterpene-1, Monoterpene-alcohol, bergamotene, bisabolane-1, β-caryophyllene, γ-elemene, sesquiterpene-1, (E)-β-farnesene, α-humulene, cadinane-1, cadinane-2, sesquiterpene, B-eudesmene, a-guaiene, bisabolane-2, sesquiterpene-3, sesquiterpene-alcohol-1, himachalane, sesquiterpene-4, unknown-compound, Guaiane-1, eudesma-3,7(11)-diene, α-farnesene, sesquiterpene-5, Guaiane-2, sesquiterpene-6, caryophyllene-oxide, guaiol, cedrol, sesquiterpene-alcohol-2, sesquiterpene-7, cadinane, unknown-compound, sesquiterpene-alcohol-3, Guaiane-4, Eremophilane, bisabolol, selinane, guaiane-5, bulnesol
[72]	35 from Ecofibre, 48 from IPK_CAN		(12) cannabidiol, cannabidiolic acid, Delta9-tetrahydrocannabinol, Delta9-tetrahydrocannabinolic acid, cannabigerol, cannabigerolic acid, cannabidivarin, cannabidivarinic acid, tetrahydrocannabivarin, tetrahydrocannabivarinic acid, cannabinol, cannabichromene

Table 6. Tripal objects and CHADO tables utilized in the ICGRC portal

Tripal entity	cvterm	Chado table	Type (if not exclusive for the chado table)
Stock	whole plant (PO:0000003)	stock	
Germplasm Accession	Germplasm Accession (CO_010:0000044)	stock	type_id
Biological Sample	biological sample (sep:00195)	biomaterial	
Analysis	Analysis (operation:2945)	analysis	
Genome Annotation	operation:Genome Annotation	analysis	type_id analysisprop genome_annotation
Genome Assembly	operation:Genome Assembly	analysis	type_id analysisprop genome_assembly
BLAST Results	BLAST results (format:1333)	analysis	type_id analysisprop blast_analysis
InterPro Results	InterPro results (local:interpro_results)	analysis	type_id analysisprop interpro_analysis
Project	Project (NCIT:C47885)	project	
Study	Study (SIO:001066)	study	
Gene	gene (SO:0000704)	feature	type_id
mRNA	mRNA (SO:0000234)	feature	type_id
Polypeptide	polypeptide (SO:0000104)	feature	type_id
QTL	QTL (SO:0000771)	feature	type_id
Genetic Map	Genetic Map (data:1278)	featuremap	type_id featuremapprop genetic
Physical Map	Physical Map (data:1280)	featuremap	type_id featuremapprop physical

Table 7. Possible tables returned by the multi-omics API built in the ICGRC portal

Input rows	Output rows	Use case
cultivar, condition	trait	Field studies
marker, condition	trait	QTL mapping
cultivar, condition, tissue	transcripts, proteins, metabolites	Transcriptomics, proteomics, metabolomics studies
marker, tissue, condition	trait	GWAS
marker, tissue, condition	expression	eGWAS, eQTL
marker, tissue, condition	metabolites	mGWAS
transcripts	metabolites	Gene/metabolite association networks,
metabolites	traits	Multivariate studies

Table 8. Definition of some Chado tables and Tripal BioEntities with ambiguous usage

Chado table	Tripal Entity	Definition
Stock		Any **stock** can be globally identified by the combination of organism, uniquename and stock type. A **stock** is the physical entities, either living or preserved, held by collections. Stocks belong to a collection; they have IDs, type, organism, description and may have a genotype.
Biomaterial		A **biomaterial** represents the MAGE concept of BioSource, BioSample, and LabeledExtract. It is essentially some **biological material** (tissue, cells, serum) that may have been processed. Processed biomaterials should be traceable back to raw biomaterials via the biomaterialrelationship table.
Organism		The organismal taxonomic classification
	Genome assembly	Genome assembly is a type of **Analysis**
Study		**Study** represents an experiment, published or otherwise, that produced microarray data
Project		**Project** standard Chado flexible property table for projects.
Analysis		An A**nalysis** is a particular type of a computational analysis; it may be a blast of one sequence against another, or an all by all blast, or a different kind of analysis altogether. It is a single unit of computation.
Assay		An A**ssay** consists of a physical instance of an array, combined with the conditions used to create the array (protocols, technician information). The assay can be thought of as a hybridization.

FIGURES

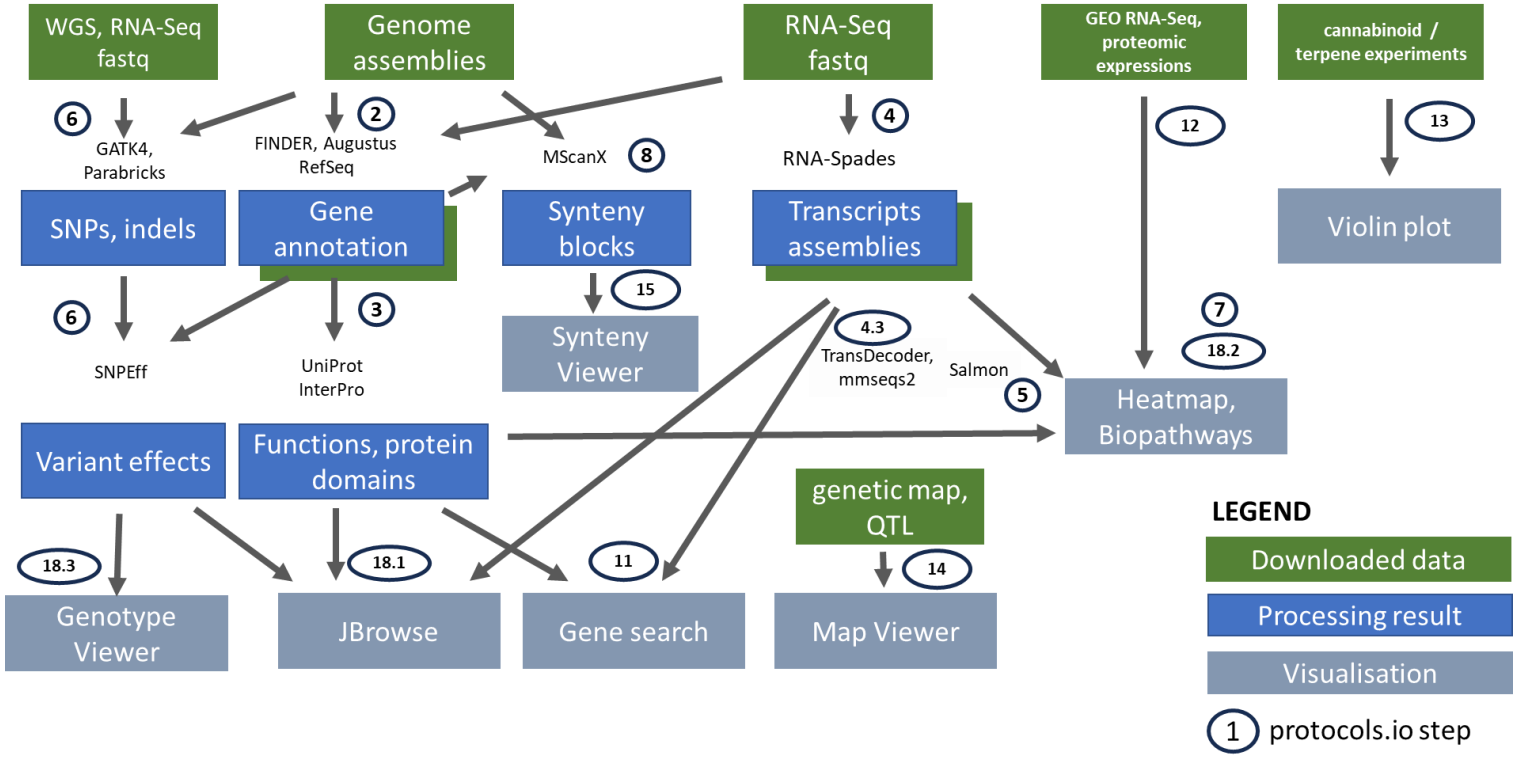

Figure 1 **Workflow for the preparation of publicly available and reanalyzed data, for the analysis tools, and visualization features in the ICGRC Cannabis Web Portal (icgrc.info).** *Downloaded data' are datasets downloaded either from public databases, supplementary files, or manually curated from articles. 'Processing result' are the new datasets generated by this project, and value added from the legacy results of the publications. 'Visualization' is the Tripal modules, and specialized web-applications included in the portal for interactive user-interface. Implementation details are documented in (protocols.io DOI: 10.17504/protocols.io.n2bvj3nz5lk5/v3).*

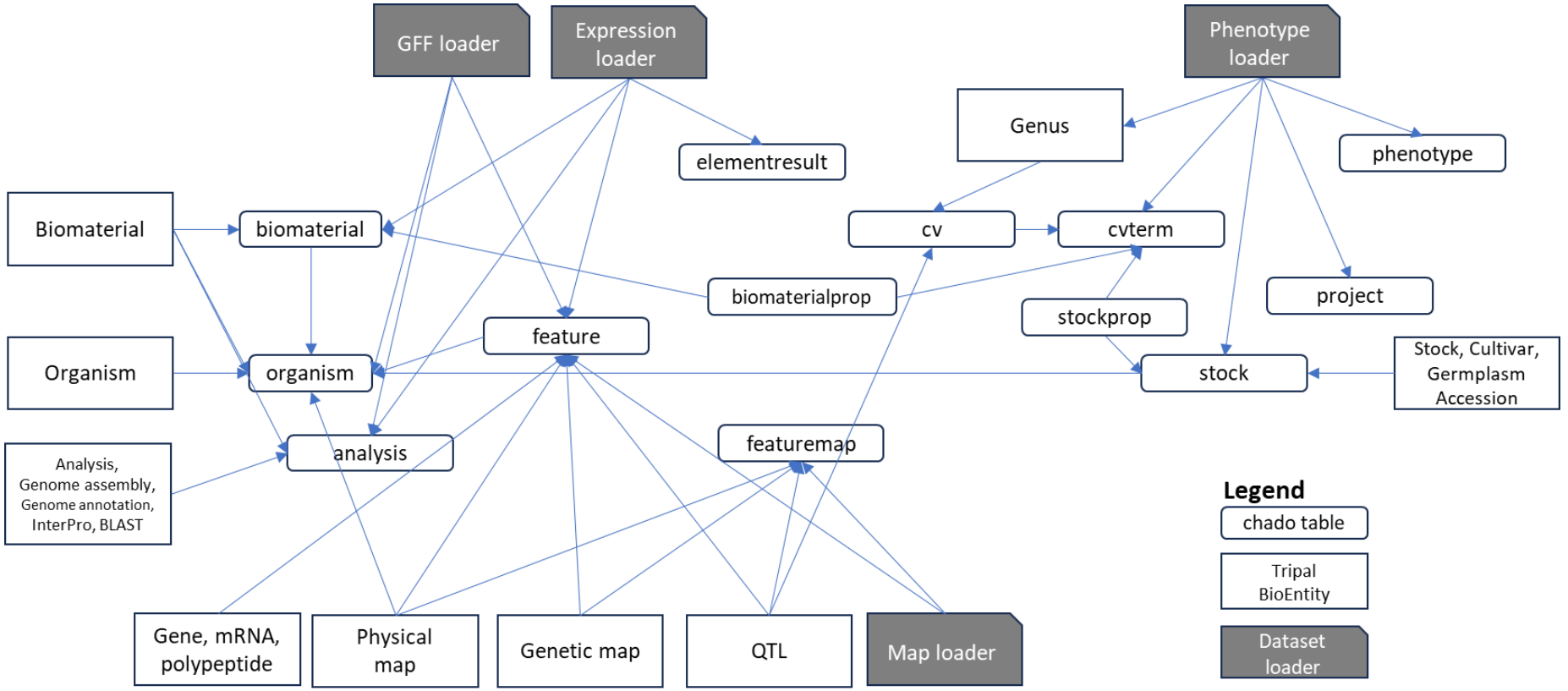

Figure 2 **Mapping of Tripal BioEntities and Chado tables.** *Input files are loaded using 'Dataset loader' for the genomic features GFF, expression, phenotype, or map datasets. Input files are loaded using 'Dataset loader' for the genomic features GFF, expression, phenotype, or map datasets. The fields are stored in various tables in the Chado schema (green). Tripal then defines BioEntities (blue) based on the record types (type_id) in the tables. The type_id's are biological ontology terms loaded in the cvterm and cv tables.*

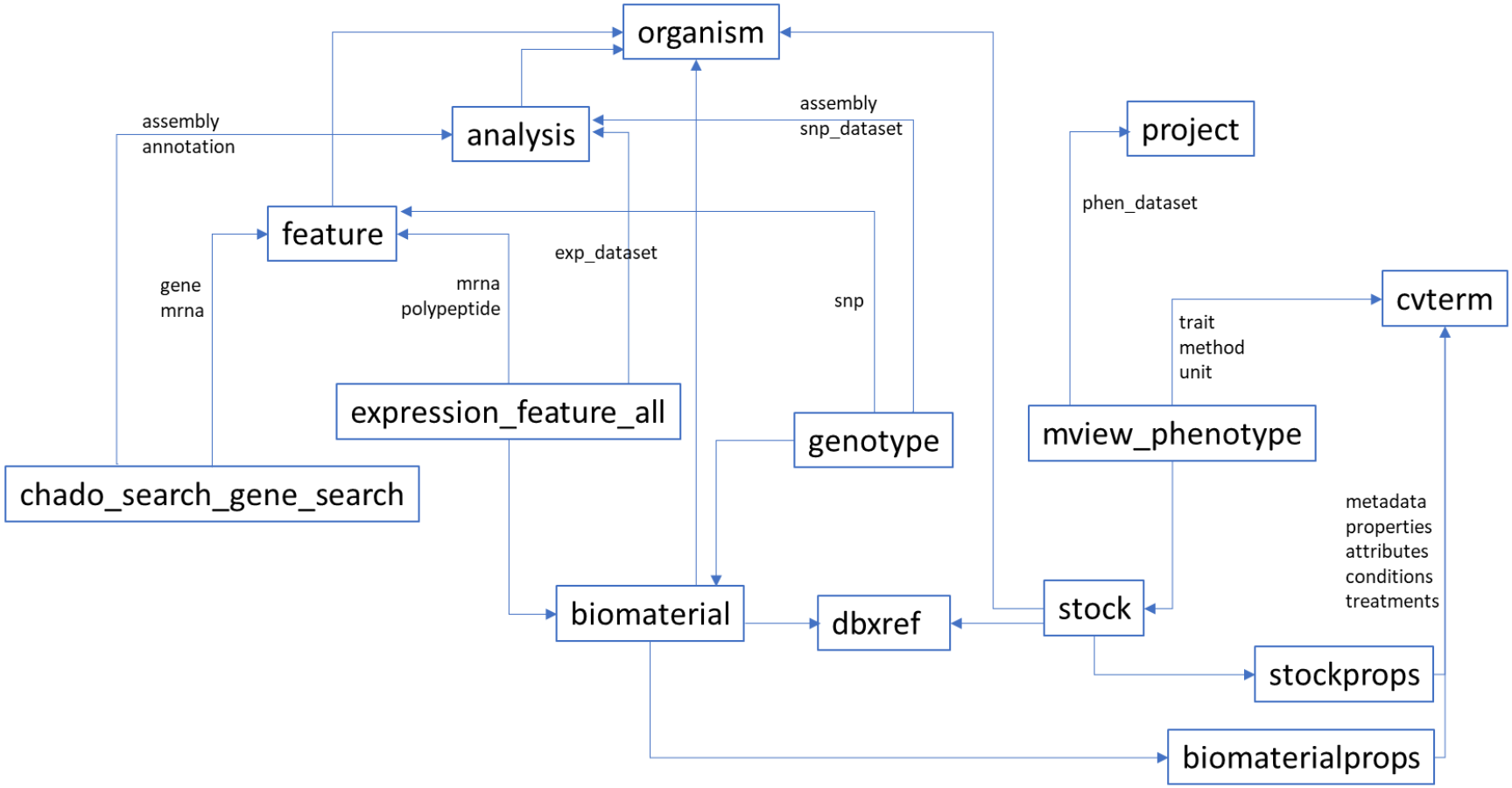

Figure 3 **Entity relationship diagram of Chado tables and materialized views involved in -omics integration.** *The Tripal modules generate materialized views to aggregate data for genomic features and gene annotations, expression levels, phenotype values. Genotype metadata on biomaterials and features are also in the database while the allele values are in external vcf.gz files. Biomaterial and Stock are synchronized through dbxref using NCBI SRR or SAMN accessions when available.*

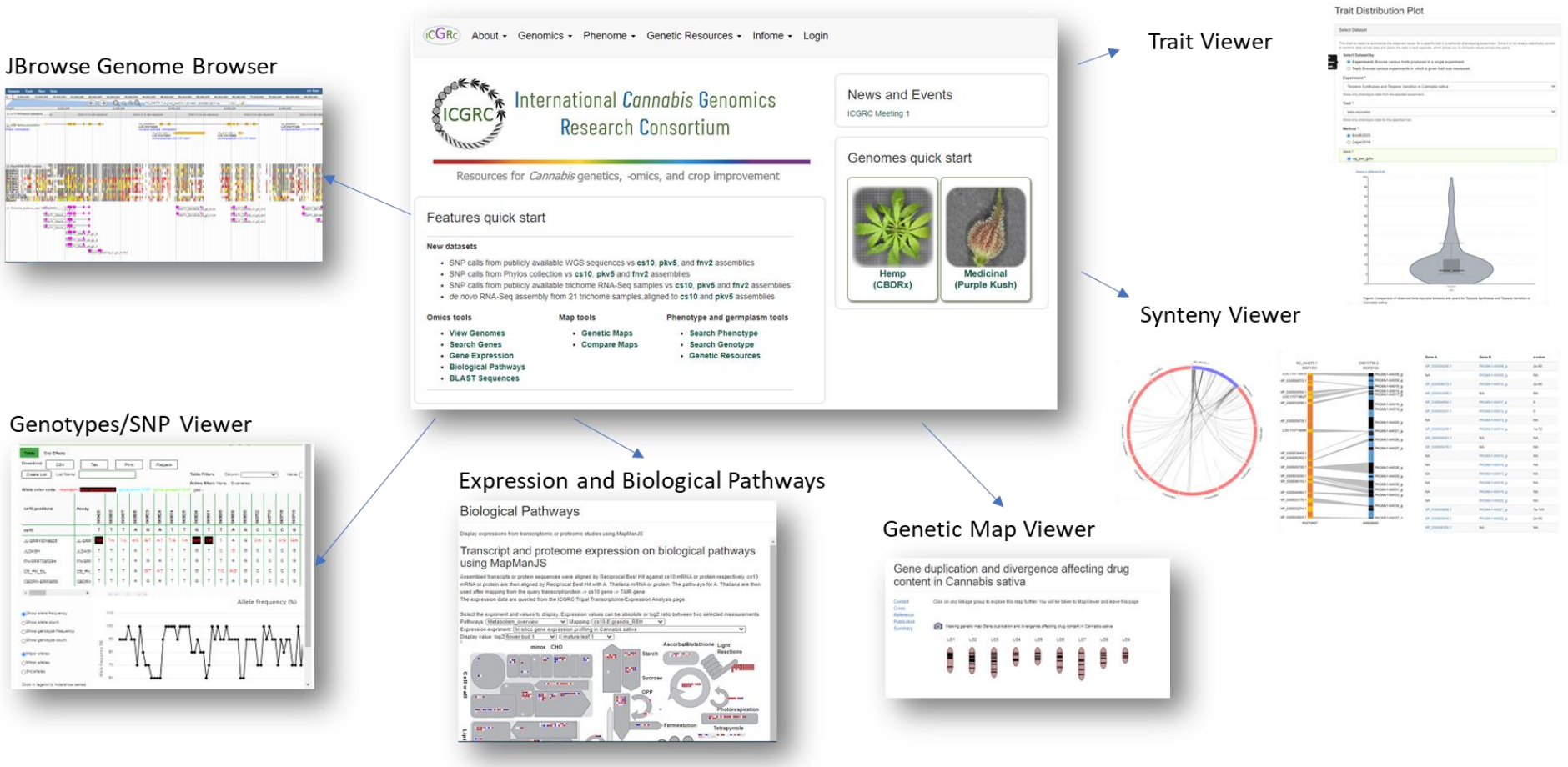

Figure 4 **ICGRC Cannabis Web Portal (icgrc.info) main feature pages.** *The web portal is a Tripal instance with various modules to host and visualize published Cannabis -omics datasets.*

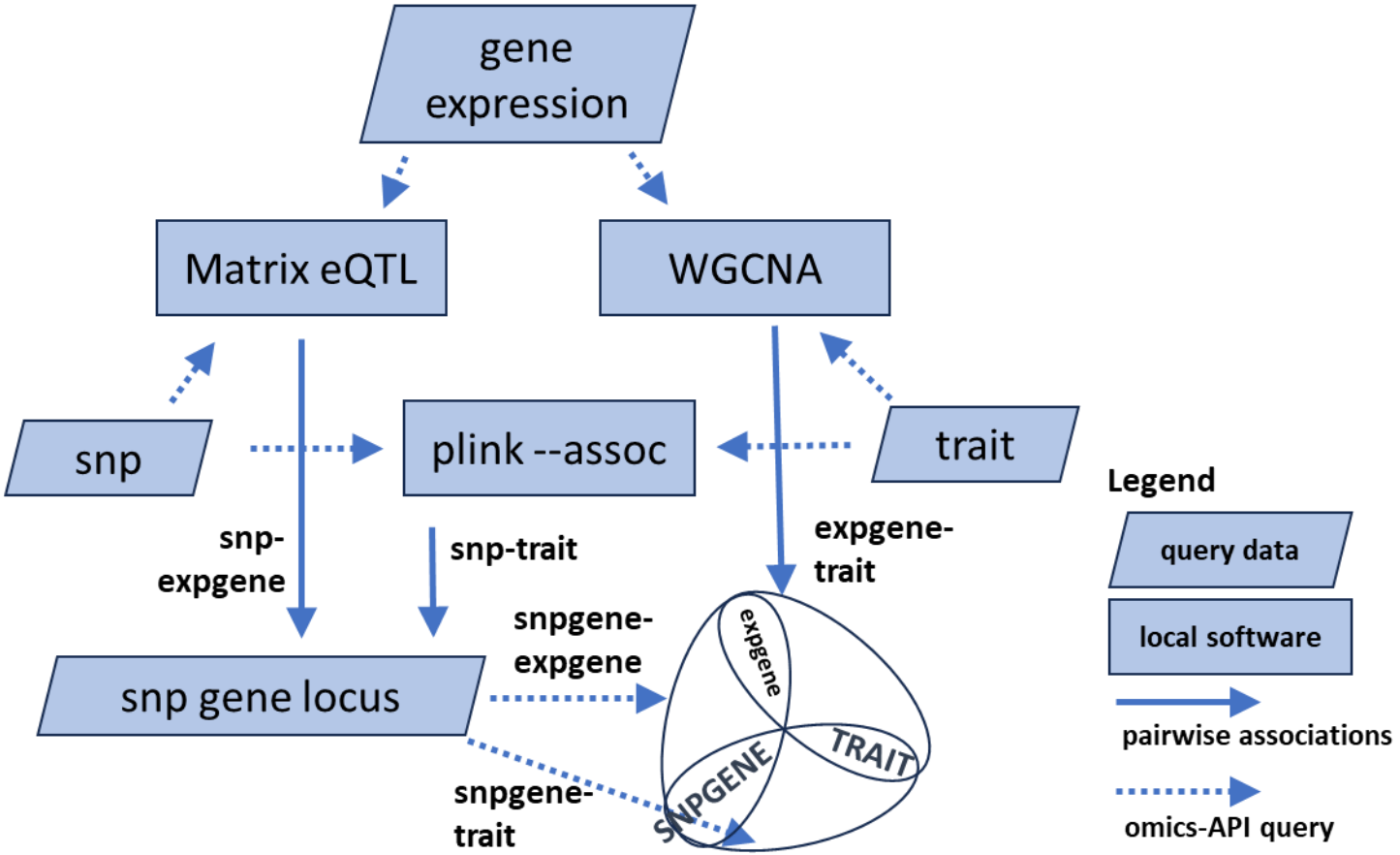

Figure 5 **Use cases on multi-omics analysis using local software and the ICGRC omics API.** *The different use-cases in the Jupyter notebook demonstration at* [99] *(https://snp.icgrc.info/static/icgrc_omics_demo.html). Results from the analysis are the top associations of SNP-transcript (Matrix eQTL), transcript-trait (WGCNA), or SNP-trait (plink -assoc) . The common datatypes between tools are SNPs (Matrix eQTL-plink), transcripts (Matrix eQTL-WGCNA), or traits (plink-WGCNA). Combining their results generates list of candidate genes with two independent evidence of associations.*

REFERENCES

1. Clarke M. Cannabis: Evolution and Ethnobiology.

2. 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*. Elsevier B.V.; 2012; doi: 10.1016/j.indcrop.2011.09.015.

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; doi: 10.3390/nu12071935.

4. Touw M. The religious and medicinal uses of Cannabis in China, India and Tibet. *J Psychoactive Drugs*. 1981; doi: 10.1080/02791072.1981.10471447.

5. Sirangelo TM, Ludlow RA. Resistance in Cannabis sativa. 2023;

6. Park SH, Staples SK, Gostin EL, Smith JP, Vigil JJ, Seifried D, et al.. Contrasting Roles of Cannabidiol as an Insecticide and Rescuing Agent for Ethanol–induced Death in the Tobacco Hornworm Manduca sexta. *Sci Rep*. 2019; doi: 10.1038/s41598-019-47017-7.

7. Desaulniers Brousseau V, Wu B Sen, MacPherson S, Morello V, Lefsrud M. Cannabinoids and Terpenes: How Production of Photo-Protectants Can Be Manipulated to Enhance Cannabis sativa L. Phytochemistry. *Front Plant Sci*. 2021; doi: 10.3389/fpls.2021.620021.

8. Fordjour E, Manful CF, Sey AA, Javed R, Pham TH, Thomas R, et al.. Cannabis: a multifaceted plant with endless potentials. *Front Pharmacol*. 2023; doi: 10.3389/fphar.2023.1200269.

9. 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*. Elsevier B.V.; 2018; doi: 10.1016/j.jpba.2017.11.073.

10. Sakamoto K, Akiyama Y, Fukui K, Kamada H, Satoh S. Characterization; Genome sizes and morphology of sex chromosomes in hemp (Cannabis sativa L.). *Cytologia (Tokyo)*. 1998; doi: 10.1508/cytologia.63.459.

11. 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; doi: 10.1186/gb-2011-12-10-r102.

12. 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; doi: 10.1111/nph.17243.

13. Expert Committee on Drug Dependence: WHO scheduling recommendations on cannabis and cannabis-related substances. https://www.who.int/publications/m/item/ecdd-41-cannabis-recommendations (2020). Accessed 2024 Feb 12.

14. businesswire: Global Cannabis Market (2021 to 2030). https://www.businesswire.com/news/home/20220203005879/en/Global-Cannabis-Market-Size-Forecast-Report-2021-A-176-Billion-by-2030---Growing-Legalization-of-Medical-Cannabis-in-Various-Countries-Driving-Growth---ResearchAndMarkets.com (2022).

15. Research and Markets: Cannabis Cultivation Market. https://www.researchandmarkets.com/reports/5165371/cannabis-cultivation-market-size-share-and-trends?utm_source=BW&utm_medium=PressRelease&utm_code=k7w29z&utm_campaign=1837122+-+Global+Cannabis+Cultivation+Market+Analysis+Report+2023%3A+A+%241%2C844+Billi (2023). Accessed 2024 Feb 12.

16. 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; doi: 10.1101/gr.242594.118.

17. Gao S, Wang B, Xie S, Xu X, Zhang J, Pei L, et al.. A high-quality reference genome of wild Cannabis sativa. *Hortic Res*. 2020; doi: 10.1038/s41438-020-0295-3.

18. Braich S, Baillie RC, Spangenberg GC, Cogan NOI. A new and improved genome sequence of Cannabis sativa. *bioRxiv*. 2020; doi: 10.1101/2020.12.13.422592.

19. 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; doi: 10.1101/2020.01.03.894428.

20. Nelson ADL, Haug-Baltzell AK, Davey S, Gregory BD, Lyons E. EPIC-CoGe: Managing and analyzing genomic data. *Bioinformatics*. 2018; doi: 10.1093/bioinformatics/bty106.

21. Tanizawa Y, Fujisawa T, Kodama Y, Kosuge T, Mashima J, Tanjo T, et al.. DNA Data Bank of Japan (DDBJ) update report 2022. *Nucleic Acids Res*. Oxford University Press; 2023; doi: 10.1093/nar/gkac1083.

22. Burgin J, Ahamed A, Cummins C, Devraj R, Gueye K, Gupta D, et al.. The European Nucleotide Archive in 2022. *Nucleic Acids Res*. 2023; doi: 10.1093/nar/gkac1051.

23. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al.. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res*. 2012; doi: 10.1093/nar/gkr944.

24. Bolser D., Staines D.M., Pritchard E. KP. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. *Plant Bioinformatics Methods Mol Biol*. p. 115–40.

25. Cao Y, She J, Li Z, Liu Y, Tian T, You Q, et al.. TomAP: A Multi-omics Data Analysis Platform for Advancing Functional Genomics Research in Tomatoes. *New Crop*. Elsevier; 2023; doi: 10.1016/j.ncrops.2023.10.001.

26. Hawkins C, Ginzburg D, Zhao K, Dwyer W, Xue B, Xu A, et al.. Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae. *J Integr Plant Biol*. John Wiley & Sons, Ltd; 2021; doi: https://doi.org/10.1111/jipb.13163.

27. Medicinal Genomics: Kannapedia. https://www.kannapedia.net (2024). Accessed 2023 Dec 1.

28. orsburnlab.org: CannabisDraftMap.org. https://www.cannabisdraftmap.org (2019). Accessed 2023 Dec 1.

29. Hughes Lab: Cannabis Genome Browser. http://genome.ccbr.utoronto.ca/index.html?org=C.+sativa&db=canSat3&hgsid=245596 (2011). Accessed 2023 Dec 1.

30. Leafly LLC: Leafly. https://www.leafly.com (2024).

31. : SeedFinder. https://en.seedfinder.eu/ (2024). Accessed 2024 Feb 12.

32. Cai. CannabisGDB: a comprehensive genomic database for Cannabis Sativa L. *Plant Biotechnol J*. :10.1111/pbi.135482021;

33. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al.. JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol*. Genome Biology; 2016; doi: 10.1186/s13059-016-0924-1.

34. Welling MT, Shapter T, Rose TJ, Liu L, Stanger R, King GJ. A Belated Green Revolution for Cannabis: Virtual Genetic Resources to Fast-Track Cultivar Development. *Front Plant Sci*. 2016; doi: 10.3389/fpls.2016.01113.

35. Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, et al.. InterMine: A flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*. 2012; doi: 10.1093/bioinformatics/bts577.

36. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, Midford PE, et al.. The MetaCyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic Acids Res*. Oxford University Press; 2020; doi: 10.1093/nar/gkz862.

37. Ficklin SP, Sanderson LA, Cheng CH, Staton ME, Lee T, Cho IH, et al.. Tripal: A construction toolkit for online genome databases. *Database*. 2011; doi: 10.1093/database/bar044.

38. Mungall CJ, Emmert DB, Gelbart WM, de Grey A, Letovsky S, Lewis SE, et al.. A Chado case study: An ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*. 2007; doi: 10.1093/bioinformatics/btm189.

39. GMOD: Generic Model Organism Database. http://gmod.org

40. Tripal.info: Tripal sites. https://tripal.info/sites_using_tripal (2008). Accessed 2024 Feb 1.

41. Staton M, Cannon E, Sanderson LA, Wegrzyn J, Anderson T, Buehler S, et al.. Tripal, a community update after 10 years of supporting open source, standards-based genetic, genomic and breeding databases. *Brief Bioinform*. 2021; doi: 10.1093/bib/bbab238.

42. southgreen.fr: SouthGreen Bioinformatics Platform. https://www.southgreen.fr/genomehubs (2013).

43. Mansueto L: ICGRC Web Portal Data Generation Protocol. protocols.io. https://www.protocols.io/private/34114B4120C511EF83720A58A9FEAC02 (2024).

44. Katz K, Shutov O, Lapoint R, Kimelman M, Rodney Brister J, O'Sullivan C. The Sequence Read Archive: A decade more of explosive growth. *Nucleic Acids Res*. Oxford University Press; 2022; doi: 10.1093/nar/gkab1053.

45. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al.. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016; doi: 10.1093/nar/gkv1189.

46. Clough E, Barrett T. The Gene Expression Omnibus database. *Methods Mol Biol*. 2016; doi: 10.1007/978-1-4939-3578-9_5.

47. Consortium TU. UniProt: the Universal Protein Knowledgebase in 2023 - Google Scholar. 51:523–312023;

48. Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, et al.. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database*. 2021; doi: 10.1093/database/baab069.

49. : Crop Ontology. https://cropontology.org/ (2008). Accessed 2023 Dec 1.

50. Banerjee S, Bhandary P, Woodhouse M, Sen TZ, Wise RP, Andorf CM. FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics*. BioMed Central; 2021; doi: 10.1186/s12859-021-04120-9.

51. 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; doi: 10.1038/s41598-019-53023-6.

52. 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; doi: 10.1111/tpj.14516.

53. Zager JJ, Lange I, Srividya N, Smith A, Markus Lange B. Gene networks underlying cannabinoid and terpenoid accumulation in cannabis. *Plant Physiol*. 2019; doi: 10.1104/pp.18.01506.

54. Booth JK, Page JE, Bohlmann J. Terpene synthases from Cannabis sativa. *PLoS One*. 2017; doi: 10.1371/journal.pone.0173911.

55. Allen KD, McKernan K, Pauli C, Roe J, Torres A, Gaudino R. Genomic characterization of the complete terpene synthase gene family from Cannabis sativa. *PLoS One*. 2019; doi: 10.1371/journal.pone.0222363.

56. McGarvey P, Huang J, McCoy M, Orvis J, Katsir Y, Lotringer N, et al.. De novo assembly and annotation of transcriptomes from two cultivars of Cannabis sativa with different cannabinoid profiles. *Gene*. Elsevier LTD; 2020; doi: 10.1016/j.gene.2020.145026.

57. Guerriero G, Behr M, Legay S, Mangeot-Peter L, Zorzan S, Ghoniem M, et al.. Transcriptomic profiling of hemp bast fibres at different developmental stages. *Sci Rep*. 2017; doi: 10.1038/s41598-017-05200-8.

58. Booth JK, Yuen MMS, Jancsik S, Madilao LL, Page AJE. Terpene synthases and terpene variation in cannabis sativa1[OPEN]. *Plant Physiol*. 2020; doi: 10.1104/PP.20.00593.

59. Massimino L. In silico gene expression profiling in Cannabis sativa. *F1000Research*. 2017; doi: 10.12688/f1000research.10631.1.

60. Gao C, Cheng C, Zhao L, Yu Y, Tang Q, Xin P, et al.. Genome-Wide Expression Profiles of Hemp (Cannabis sativa L.) in Response to Drought Stress. *Int J Genomics*. Hindawi; 2018; doi: 10.1155/2018/3057272.

61. Conneely LJ, Mauleon R, Mieog J, Barkla BJ, Kretzschmar T. Characterization of the Cannabis sativa glandular trichome proteome. *PLoS One*. 2021; doi: 10.1371/journal.pone.0242633.

62. ICGRC CannSeek: ICGRC CannSeek. https://icgrc.info/genotype_viewer (2022).

63. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al.. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; doi: 10.1101/gr.107524.110.20.

64. NVIDIA: Parabricks. https://docs.nvidia.com/clara/parabricks/latest/index.html (2023). Accessed 2023 Dec 1.

65. Mansueto L, Fuentes RR, Borja FN, Detras J, Abrio-Santos JM, Chebotarov D, et al.. Rice SNP-seek database update: New SNPs, indels, and queries. *Nucleic Acids Res*. 2017; doi: 10.1093/nar/gkw1135.

66. Podman: Podman. https://podman.io/ (2024).

67. Docker: Docker. https://www.docker.com

68. Buble K, Jung S, Humann JL, Yu J, Cheng CH, Lee T, et al.. Tripal MapViewer: A tool for interactive visualization and comparison of genetic maps. *Database*. 2019; doi: 10.1093/database/baz100.

69. Sanderson LA, Caron CT, Tan R, Shen Y, Liu R, Bett KE. KnowPulse: A Web-Resource Focused on Diversity Data for Pulse Crop Improvement. *Front Plant Sci*. 2019; doi: 10.3389/fpls.2019.00965.

70. Welling MT, Liu L, Shapter T, Raymond CA, King GJ. Characterisation of cannabinoid composition in a diverse Cannabis sativa L. germplasm collection. *Euphytica*. Springer Netherlands; 2016; doi: 10.1007/s10681-015-1585-y.

71. Richins RD, Rodriguez-Uribe L, Lowe K, Ferral R, O'Connell MA. Accumulation of bioactive metabolites in cultivated medical Cannabis. *PLoS One*. 2018; doi: 10.1371/journal.pone.0201119.

72. Gloerfelt-Tarp F, Hewavitharana AK, Mieog J, Palmer WM, Fraser F, Ansari O, et al.. Using a global diversity panel of Cannabis sativa L. to develop a near InfraRed-based chemometric application for cannabinoid quantification. *Sci Rep*. Nature Publishing Group UK; 2023; doi: 10.1038/s41598-023-29148-0.

73. Degtyarenko K, De matos P, Ennis M, Hastings J, Zbinden M, Mcnaught A, et al.. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res*. 2008; doi: 10.1093/nar/gkm791.

74. Laporte M-A, elserj, Mungall C, Cooper L, ejohnson1, Hoyt CT, et al.. Planteome/plant-experimental-conditions-ontology: Updating PECO files for Planteome release 5. Zenodo;

75. Gkoutos G V., Schofield PN, Hoehndorf R. The Units Ontology: a tool for integrating units of measurement in science. *Database (Oxford)*. 2012; doi: 10.1093/database/bas033.

76. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; doi: 10.1093/bioinformatics/btu170.

77. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Research*. 2020; doi: 10.12688/f1000research.23297.2.

78. Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*. 2016; doi: 10.1093/bioinformatics/btw006.

79. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al.. InterProScan: Protein domains identifier. *Nucleic Acids Res*. 2005; doi: 10.1093/nar/gki442.

80. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. RnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience*. Oxford University Press; 2019; doi: 10.1093/gigascience/giz100.

81. Wu TD, Watanabe CK. GMAP: A genomic mapping and alignment program for mRNA and EST

sequences. *Bioinformatics*. 2005; doi: 10.1093/bioinformatics/bti310.

82. Haas B: Transdecoder. https://github.com/TransDecoder/TransDecoder (2012).

83. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al.. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012; doi: 10.1093/nar/gkr1293.

84. Schwacke R, Ponce-Soto GY, Krause K, Bolger AM, Arsova B, Hallab A, et al.. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Mol Plant*. 2019; doi: 10.1016/j.molp.2019.01.003.

85. Broad Institute: GATK Germline Pipeline. https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels (2018).

86. Mansueto L: ICGRC Portal Tripal Data Generation and Setup. https://www.protocols.io/view/icgrc-portal-tripal-data-generation-and-setup-n2bvj3nz5lk5 (2023).

87. Nadkarni PM, Marenco L, Chen R, Skoufos E, Shepherd G, Miller P. Organization of heterogeneous scientific data using the EAV/CR representation. *J Am Med Informatics Assoc*. 1999; doi: 10.1136/jamia.1999.0060478.

88. Sanderson LA, Caron CT, Tan RL, Bett KE. A PostgreSQL Tripal solution for large-scale genotypic and phenotypic data. *Database*. 2021; doi: 10.1093/database/baab051.

89. Danecek P, McCarthy SA. BCFtools/csq: Haplotype-aware variant consequences. *Bioinformatics*. 2017; doi: 10.1093/bioinformatics/btx100.

90. Selby P, Abbeloos R, Backlund JE, Basterrechea Salido M, Bauchet G, Benites-Alfaro OE, et al.. BrAPI - An application programming interface for plant breeding applications. *Bioinformatics*. 2019; doi: 10.1093/bioinformatics/btz190.

91. Tripal: Tripal Web Services. https://tripal.readthedocs.io/en/latest/user_guide/web_services.html

92. ICGRC: ICGRC Tripal Web Services. https://icgrc.info/web-services/content/v0.1

93. ICGRC: ICGRC Portal. https://icgrc.info (2022).

94. ICGRC API: ICGRC Omics API Documentation. https://icgrc.info/api_doc (2022). Accessed 2024 Feb 12.

95. Manzella D, Marsella M, Jaiswal P, Arnaud E, King B. Digital Sequence Information and Plant Genetic Resources: Global Policy Meets Interoperability. *Towar Responsible Plant Data Link Data Challenges Agric Res Dev*. 2023; doi: 10.1007/978-3-031-13276-6_10.

96. Torkamaneh D, Jones AMP. Cannabis, the multibillion dollar plant that no genebank wanted. *Genome*. 2022; doi: 10.1139/gen-2021-0016.

97. CropTrust: Genesys. https://www.genesys-pgr.org/ (2017).

98. Andrés-Hernández L, Halimi RA, Mauleon R, Mayes S, Baten A, King GJ. Challenges for FAIR-compliant description and comparison of crop phenotype data with standardized controlled vocabularies. *Database*. 2021; doi: 10.1093/database/baab028.

99. ICGRC API Demo: ICGRC Omics API Demo. https://snp.icgrc.info/static/icgrc_omics_demo.html

(2023). Accessed 2024 Feb 12.

100. : scikit-learn. https://scikit-learn.org/ (2011). Accessed 2023 Nov 1.

101. : pyppca. https://github.com/shergreen/pyppca (2019). Accessed 2023 Nov 1.

102. Behdenna A, Haziza J, Azencott C-A, Nordor A. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *bioRxiv*. :2020.03.17.9954312021;

103. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008; doi: 10.1186/1471-2105-9-559.

104. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012; doi: 10.1093/bioinformatics/bts163.

105. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al.. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; doi: 10.1086/519795.