DIPARTIMENTO DI SCIENZE AGRARIE ALIMENTARI E AMBIENTALI

CORSO DI LAUREA IN: SCIENZE AGRARIE E DEL TERRITORIO

CURRICULUM: PRODUZIONE E PROTEZIONE DELLE COLTURE VEGETALI

IMPIEGO DELL'INTELLIGENZA ARTIFICIALE PER LA CARATTERIZZAZIONE DELLE RISORSE GENETICHE IN PROGETTI DI CONSERVAZIONE DECENTRALIZZATA

Use of artificial intelligence for characterization of genetic resources in decentralized conservation projects

TESI SPERIMENTALE

Studente:
Giacomo Paolinelli

Relatore:
PROF. ROBERTO PAPA

Correlatore:
DR. VALERIO DI VITTORI
PROF. ADRIANO MANCINI

ANNO ACCADEMICO 2023-2024

Alla mia famiglia
A Veronica

SOMMARIO

ELENCO DELLE TABELLE

ELENCO DELLE FIGURE

ACRONIMI E ABBREVIAZIONI

CBD	Convenzione sulla diversità biologica
ISPRA	Istituto Superiore per la Protezione e la Ricerca Ambientale
FAO	Food and Agricolture Organization of the United Nations
PGRs	Plant Genetic Resources
ITP-GRFA	International Treaty on Plant Genetic Resources for Food and Agricolture
DNA	Acido desossiribonucleico
EURISCO	European Search Catalogue for Plant Genetic Resources
CGN	Center of Genetic Resources
IPGRI	International Plant Genetic Resources institute
EPGRIS	European Plant Genetic Resources Information Infrastructure
NGB	Nordic GeneBank
IPK	The Leibniz Institute of Plant Genetics and Crop plant Research
WGS	Whole Genome Sequencing
GBS	Genotyping by Sequencing
RNA	Acido Ribonucleico
CNR	Consiglio Nazionale delle Ricerche
QTL	Quantitative Trait Locus
SSD	Single Seed Descent
DOI	Digital Object Identifier
RGB	Red; Green; Blue
mRNA	RNA messaggero
GWAS	Genome-Wide Association Studies (GWAS)

DArT	Diversity Array Technology
SSRs	Simple Sequence Repeats
SNPs	Single Nucleotide Polymorphisms
NGS	Next Generation Sequencing
IPCC	Intergovernmental Panel on Climate Change
N2	Azoto
NH3	Ammoniaca
INCREASE	Intelligent Collections of Food Legumes Resources for European Agrofood System
M1	Jalisco e Durango race
M2	Mesoamerica race
A1	Nueva Granada race
A2	Perù
A3	Chile
MBP	Mega base pair
Gbp	Giga base pair
ICs	Intelligent Collections
R-CORE	Reference Core
T-CORE	Training core
H-CORE	Hyper core
ML	Machine Learning
AI	Artificial Intelligence
CSE	Citizen Science Experiment
INCREASE CSA	INCREASE Citizen Science App
eSMTA	Standard Material Transfer Agreement
CNNs	Convolutional Neural Network
GS	Genomic Selection

GANs	Generative Adversial Networks
GDD	Growing Degree Days
GPU	Graphics Processing Unit
UNIVPM	Università politecnica delle Marche
YOLO	You Only Look Once
SOTA	State of The Art
JSON	Java Script Object Notation
.XML	Extensible Markup Language
.TXT	Text File
RUs	Randomization Unit
EU	European Union
DTF	Days to Flowering
MAP	Mean avarage precision
PS	Photoperiod Sensitivity

INTRODUZIONE

1.1 **Risorse genetiche nelle leguminose alimentari**

1.1.1 *L'importanza della Biodiversità e della sua conservazione*

La convenzione sulla diversità biologica, stipulata dalle Nazioni Unite nel 1992, definisce il termine "Biodiversità" o "Diversità Biologica" come la variabilità tra gli organismi viventi di ogni origine, dagli ecosistemi terrestri agli ecosistemi marini ed i complessi ecologici di cui essi fanno parte includendo la diversità intraspecifica, interspecifica e a livello di ecosistema (CBD, 1992). Gli obiettivi della convenzione sono quelli di garantire la conservazione della biodiversità, l'uso sostenibile delle sue componenti e l'equilibrata distribuzione dei benefici ottenuti attraverso l'impiego delle risorse genetiche (CBD, 1992), definite come "materiale genetico con un valore attuale o potenziale" (CBD, 1992) e come "qualsiasi materiale animale e/o vegetale che contiene unità funzionali di eredità" (CBD, 1992). La convenzione sottolinea poi l'importanza dell'accesso alle risorse genetiche ed il trasferimento delle conoscenze relative ad esse ed alle tecnologie impiegate (CBD, 1992). L'Istituto Superiore per la protezione e la Ricerca Ambientale (ISPRA) definisce i sottolivelli di Biodiversità (Interspecifica, Genetica ed Ecosistema), così come indicato nella Figura 1.

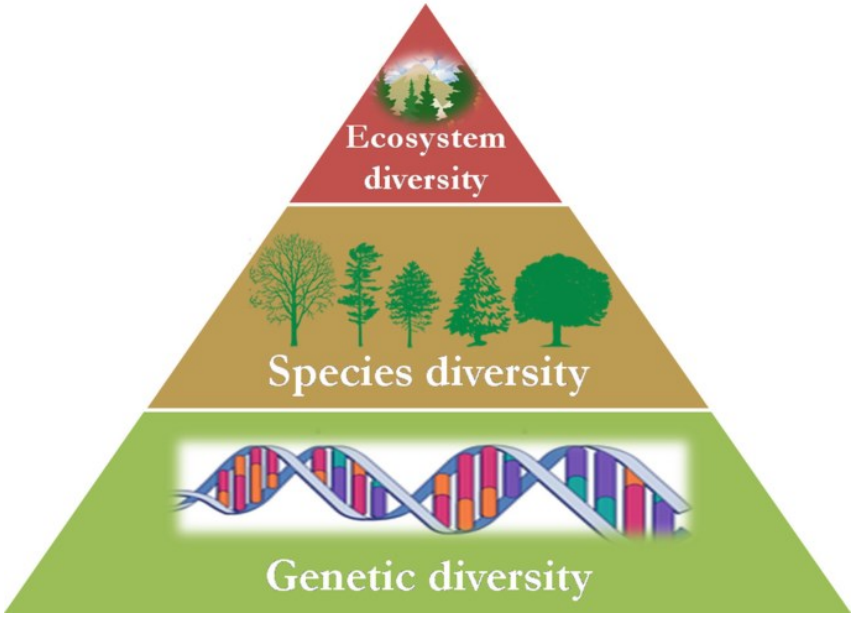


Figura 1: Classificazione dei livelli di biodiversità (ISPRA). La diversità di ecosistema è rappresentata dal numero e dall'abbondanza degli habitat, delle comunità viventi e degli ecosistemi all'interno dei quali i diversi organismi che lo popolano vivono ed evolvono (ISPRA); la Diversità di specie comprende la numerosità delle specie presenti in una data zona (ricchezza in specie) o di rarità/abbondanza in un dato territorio o habitat (frequenza delle specie) (ISPRA); la diversità genetica (intraspecifica) è caratterizzata dalla totalità del patrimonio genetico e quindi tutte le possibili differenze tra i geni all'interno di una data specie. Rappresenta quindi tutto il materiale ereditabile degli organismi che popolano la terra. Figura tratta da Nonic et al. 2021.

La biodiversità può essere misurata ed osservata in qualsiasi scala spaziale, da un micro-sito ad un habitat fino all'intera biosfera (Swingland et al. 2001). Whittaker (1972) definisce tre tipologie di diversità ovvero α – *Diversità*, cioè la ricchezza specifica totale, (numerosità delle specie presenti in un campione o in una singola comunità ecologica), β – *Diversità*, ovvero la diversità di specie tra comunità (variazione della composizione specifica lungo un gradiente ambientale o geografico) e γ – *Diversità*, ossia la diversità di specie a livello regionale (numerosità delle specie in un'area molto estesa, anche a livello di continente) (Casavecchia, 2023). Ad oggi si stima che il numero di piante superiori sia compreso tra le 300.000 e le 500.000. Di queste 250.000 sono state identificate (Wilson, 1988; Heywood, 1995) e 30.000 commestibili. Secondo i dati FAO, l'uomo utilizza per la sua alimentazione circa 7.000 specie vegetali ma ne coltiva solo 150 di queste; il 75% degli alimenti consumati dall'uomo è fornito

solo da 12 specie vegetali, dove il 50% di questo è rappresentato da 4 specie (riso, mais, grano e patata) (Ministero dell'agricoltura. 2008).

La diversità genetica è l'ammontare totale della variabilità genetica riscontrabile tra individui di una varietà, popolazioni diverse o specie (Brown 1983), ed è il risultato dell'azione congiunta di ricombinazione, processi evolutivi e di adattamento; infatti, mutazioni, flusso genico, selezione (Salgotra et al. 2023) e deriva genetica (Brown 1983) contribuiscono a plasmare la variabilità riscontrabile a livello genotipico (sequenza di DNA, profilo epigenetico), e a livello fenotipico (strutture proteiche, isoenzimi, fisiologia e morfologia) (Salgotra et al. 2023) (Figura 2).

Figura 2:Variabilità fenotipica osservabile in Landraces italiane di fagiolo. Figura tratta da Lioi et al. 2013.

La Diversità genetica è la base per la sopravvivenza delle piante in natura e per il miglioramento genetico; infatti, la variabilità delle risorse fitogenetiche permette di sviluppare nuove e migliori cultivar con le caratteristiche desiderate in programmi di miglioramento genetico (Bhandari et al. 2017).

Recentemente abbiamo assistito ad un eccessivo sfruttamento e depauperamento delle risorse genetiche a causa del cambiamento climatico, dell'introduzione di specie aliene competitive e dell'aumento demografico (Salgotra et al. 2023), unitamente all'affermazione della monocultura in campo dovuta alla sostituzione di landraces locali con varietà migliorate più performanti e redditizie (Salgotra et al. 2023). Questi aspetti hanno contribuito alla

riduzione della biodiversità agraria; infatti, tale fenomeno viene riconosciuto con il nome di "Erosione Genetica". I dati FAO mostrano che oltre il 75% delle diversità genetica nelle PGRs e il 90% delle varietà localmente coltivate sono state perse e sono scomparse dai campi degli agricoltori (FAO, 2004) (Salgotra et al. 2023).

1.1.2 *Risorse genetiche e strategie di conservazione*

Sulla base di quanto illustrato nel precedente paragrafo, possiamo affermare che la diversità genetica è una risorsa fondamentale in termini generali e per quanto riguarda l'impiego in programmi di miglioramento genetico. La diversità racchiusa nelle specie coltivate rientra nella definizione di "Plant Genetic Resources" (PGRs) (Risorse Genetiche delle piante) (Salgotra et al. 2023). Tali risorse sono parte integrante della grande biodiversità che caratterizza la vita sulla terra e come riportato precedentemente, vengono definite come "Materiale con un valore reale e potenziale" (CBD, 1992) afferente al mondo delle colture d'interesse nel settore agro-alimentare. Le risorse genetiche sono tutti quei materiali e risorse che racchiudono alleli con un potenziale interesse economico, nutrizionale ed ambientale, ed impiegabili in programmi di miglioramento genetico al fine di ottenere varietà migliorate, resistenti e ben adattate a diversi ambienti.

Esempi di PGRs sono:

- *Wild Species (Specie spontanee)*: specie che non hanno subito il processo di domesticazione. La loro utilità può essere diretta o indiretta, attuale o potenziale;
- *Wild Relatives*: rappresentano i progenitori delle specie che hanno subito la domesticazione. Sono genotipi con utilità potenziale nel miglioramento genetico;
- *Ecotipi*: sono popolazioni spontanee adattate ad un determinato ambiente indipendentemente dall'intervento umano.
- *Landraces (Varietà Locali) (Figura 2-3)*: sono popolazioni caratterizzate da un adattamento specifico ad un determinato ambiente e metodologie di coltivazione, quindi strettamente legato a usi, conoscenze, abitudini e tradizione della popolazione che l'ha sviluppata e ne mantiene viva la coltivazione;
- *Varietà Migliorate*: derivano da programmi di miglioramento genetico condotti da Breeder professionisti. Sono popolazioni omogenee, il più delle volte costituite da un solo genotipo (Linee Pure, Ibridi semplici, cloni etc.);
- *Varietà Moderne coltivate*;

- *Stocks Genetici*: collezioni di varianti genetiche utilizzate nella ricerca (Bitocchi, Tesi 2004). Tra questi possiamo riconoscere le "Breeder's Lines", "Elite Lines" e "Mutanti" (Salgotra et al. 2023);

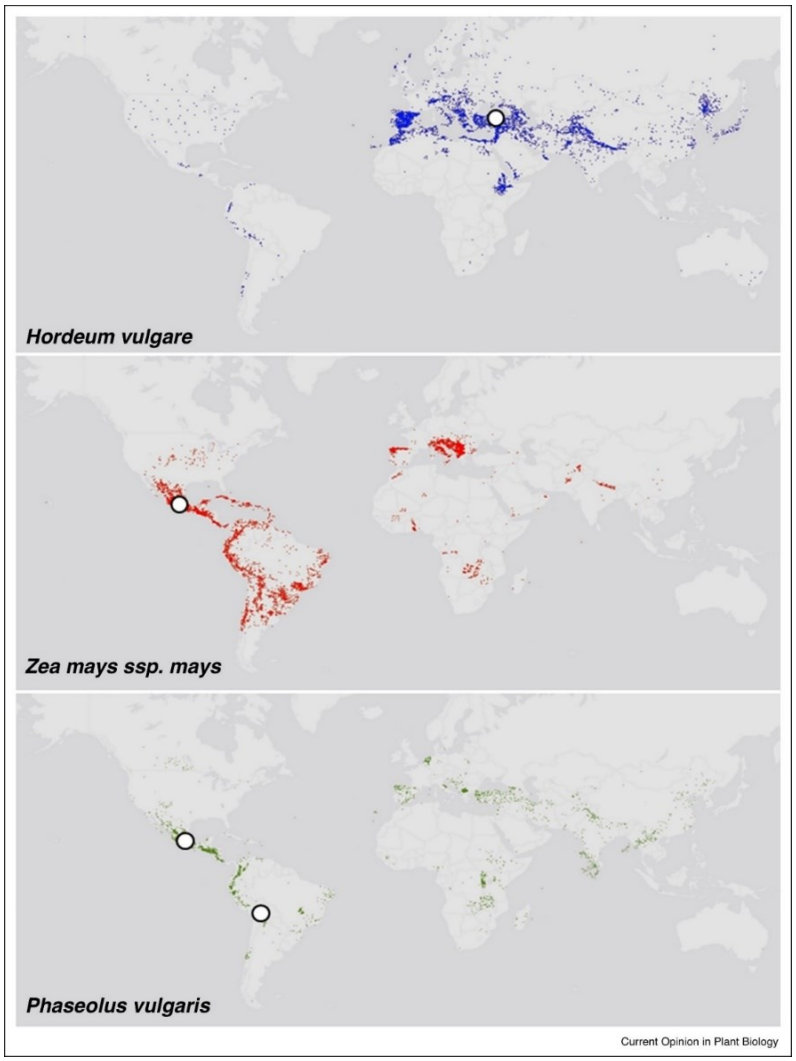


Figura 3: Distribuzione geografica di landraces di Orzo (in alto), Mais (centro) e Fagiolo comune (in basso) rispetto ai centri di domesticazione primaria (cerchio bianco con contorno nero). Figura tratta da Cortinovis et al. (2020)

Le risorse genetiche sono fondamentali nell'ambito del miglioramento genetico, della resilienza del sistema agricolo, della sicurezza alimentare, dei servizi ecosistemici (Wambugu et al. 2023) e della lotta al cambiamento climatico (Wambugu et al. 2018) (Salgotra et al. 2023). Il livello e la distribuzione della diversità genetica tra ed entro specie coltivate permette di indagare aspetti cruciali dell'evoluzione, quali l'origine, la domesticazione, la diffusione e

adattamento di una specie di interesse agrario (Bitocchi et al. 2012-2013, Cortinovis et al. 2020, Bellucci et al. 2023). Il sequenziamento dei genomi di specie di interesse agrario (Schmutz et al., 2014, The Tomato Genome Consortium. 2012; International Rice Genome Sequencing Project 2005), lo sviluppo di pan-genomi (Liu et al. 2020, Song et al. 2020, Gao et al. 2019, Cortinovis et al. 2023), e studi di genomica di popolazione e di paesaggio (e.g., landscape genomics) contribuiscono a definire al livello molecolare la complessità dell'interazione con l'ambiente e l'evoluzione di una specie (Cortinovis et al. 2020).

Con riferimento alle problematiche evidenziate nel precedente paragrafo (i.e., erosione genetica), sono stati sviluppati diversi accordi internazionali in armonia con la "Convenzione sulla Diversità Biologica" (CBD) per la conservazione, utilizzo sostenibile, equità nella ridistribuzione dei benefici e l'utilizzo sicuro delle risorse genetiche (Salgotra et al. 2023). Tra questi trattati, oltre alla CBD, tra i più importanti riconosciamo, "*The Cartagena Protocol on Biosafety*" (2000), "*The International Treaty on Plant Genetic Resources for Food and Agricolture*" (ITP-GRFA) , ed il "*Protocollo di Nagoya*" (2014) (Salgotra et al. 2023).

La conservazione delle risorse genetiche si avvale di diversi approcci in funzione della biologia riproduttiva e la natura degli organi di propagazione della specie in questione, ed altri aspetti tecnici come la disponibilità di risorse umane, finanziarie ed istituzionali per operare una conservazione efficiente nel lungo tempo (Lorenzetti et al. 2021). La conservazione delle risorse genetiche può essere di tipo statico oppure dinamico (Figura 4); nel primo caso il materiale genetico viene collezionato e mantenuto in condizione di staticità (e.g., conservazione dei semi all'interno di una banca del germoplasma), cioè senza pressioni evolutive, determinando un invariato profilo genetico nel tempo. Nella conservazione dinamica, invece, il materiale è soggetto a pressione evolutiva (e.g., conservazione dinamica in campo per più generazioni) comportando invece la variazione nel tempo del proprio genoma per via dell'azione delle forze evolutive.

Le strategie per la conservazione possono essere distinte in due macrocategorie: in Situ ed ex Situ (Figura 4). La conservazione in situ, che prevede la conservazione della risorsa genetica nell'areale di origine in cui viene collezionata, è una tipologia dinamica di conservazione, mentre nel caso della conservazione ex situ, che prevede la conservazione in un areale differente da quello del campionamento originale, possiamo avere situazioni statiche o dinamiche a seconda del programma adottato.

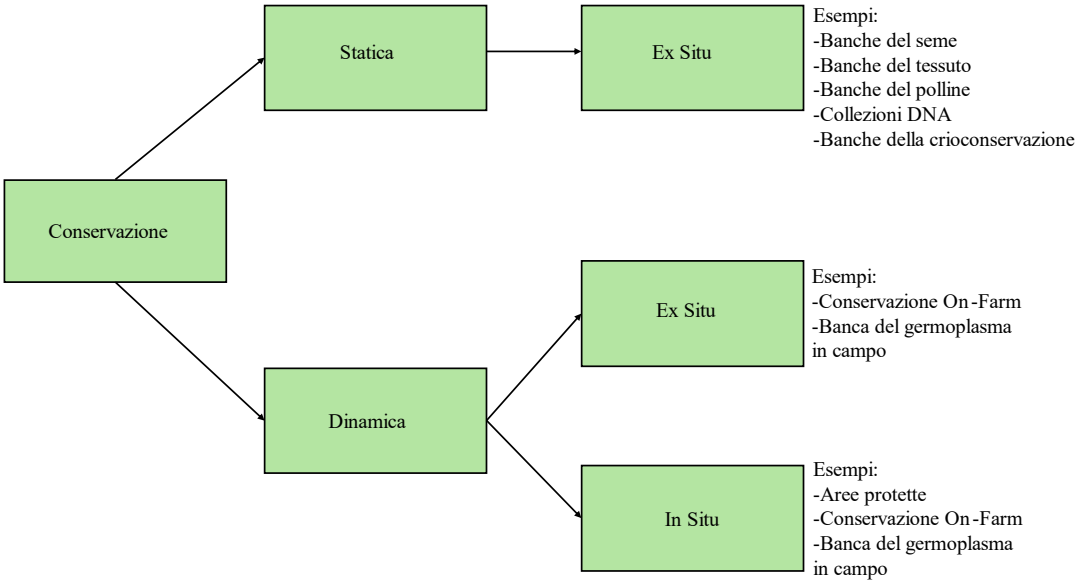


Figura 4: Strategie di conservazione per le risorse genetiche delle piante. Figura tratta da Joshi et al. (2018) con modifiche.

La conservazione in Situ delle risorse genetiche può essere applicata prevalentemente tramite due metodi (Figura 4):

- *Conservazione in azienda e/o in campo* riguarda prevalentemente le varietà locali ed i loro metodi di coltivazione, entrambi legati alla tradizione agricola di un dato luogo (Salgotra et al. 2023). In questo caso è quindi l'agricoltore stesso che si rende custode della biodiversità e permette la sopravvivenza di varietà locali (Landraces).
- *Istituzione di riserve ed aree protette per la conservazione delle risorse genetiche* concerne la salvaguardia di interi habitat naturali e quindi delle specie (prevalentemente selvatiche) che lo popolano. (Salgotra et al. 2023).

Per quanto riguarda la conservazione ex Situ invece in questo caso si tratta di conservare le risorse genetiche al di fuori dal loro habitat naturale (Figura 4). Questi metodi di conservazione possono risultare fondamentali per preservare le risorse genetiche di specie a rischio estinzione e/o materiali in pericolo (e.g., Svalbard https://www.seedvault.no/) per eventuali calamità naturali, interferenza umana, cambiamento climatico, sovrasfruttamento e sovrautilizzazione (Salgotra et al. 2023). Tra le metodologie "ex Situ" di conservazione particolare rilevanza è assunta da fondamentali strutture definite come banche del germoplasma. Il secondo report FAO (2009) sulla conservazione delle risorse genetiche riporta come in tutto il mondo siano conservate circa 7.4 milioni di accessioni (Campione distinto, univocamente identificabile di semi che rappresentano una cultivar, una "breeding line" o una

popolazione che viene conservato e utilizzato. FAO Glossary) di cui circa il 45% riguardano colture cerealicole, seguite da accessioni di specie leguminose alimentari (15%), foraggere (9%) e orticole (7%) come mostrato nella tabella 1.

Tabella 1: Germoplasma mondiale conservato ex situ dai centri di ricerca nazionali ed internazionali distinto per gruppi di specie (FAO-WIEWS. 2009)

Gruppi di specie	**Numero di accessioni**	**Specie selvatiche (%)**	**Varietà locali (*Landraces*) (%)**	**Materiale per il miglioramento (%)**	**Cultivar (%)**	**Altro (%)**
Cereali	3.157.578	5	29	15	8	43
Leguminose alimentari	1.069.897	4	32	7	9	48
Radici e tuberi	204.408	10	30	13	10	37
Orticole	502.408	5	22	8	14	51
Frutta a guscio, fruttiferi e bacche	423.401	7	13	14	21	45
Oleaginose	181.752	7	22	14	11	46
Foraggere	651.024	35	13	3	4	45
Colture da zucchero	63.474	7	7	11	25	50
Pianta da fibra	169.969	4	18	10	10	58
Piante medicinali, aromatiche e spezie	160.050	13	24	7	9	47
Colture industriali e ornamentali	152.325	46	1	2	4	47
Altro	262.993	29	4	2	2	63
Totale	6.999.279	10	24	11	9	46

In tutto il mondo, possiamo contare almeno 1800 banche del germoplasma (Genebanks) che conservano risorse genetiche vegetali, di cui circa 625 sono situate in Europa dove possiamo trovare più di 2 milioni di accessioni (Engels et al. 2012). Tra le tappe fondamentali nel processo di conservazione delle risorse genetiche, nel 2001 è stato istituito il "Catalogo Europeo di Ricerca delle Risorse Genetiche delle Piante" (EURISCO) nell'ambito di un progetto europeo EPGRIS (European Plant Genetic Resources Information Infrastructure) coordinato dal centro per le risorse genetiche "The Netherlands" (CGN), con la partecipazione della Repubblica Ceca, Francia, Germania, Portogallo, l'Istituto Internazionale delle Risorse Genetiche delle Piante (IPGRI, ora Biodiversity International) e la "Nordic Gene Bank" (NGB, ora NordGen) (Weise et al. 2017). Inoltre, nella piattaforma EURISCO sono disponibili informazioni su circa 1.8 milioni di accessioni di specie vegetali (Weise et al. 2017) (https://eurisco.ipk-gatersleben.de/apex/eurisco_ws/r/eurisco/home). Dal 2014, l'istituto di Genetica Vegetale e di Ricerca sulle Piante Coltivate "Leibniz" di Gatersleben in Germania (IPK, https://www.ipk-gatersleben.de/en/) coordina le attività di gestione e sviluppo del network EURISCO (Weise et al. 2016). L'istituto IPK (figura 5) rappresenta una delle banche del germoplasma tra le più grandi al mondo, in termini di accessioni e diversità custodita (IPK). La banca del germoplasma "The Federal Ex Situ Gene Bank" conserva attualmente 151.348 accessioni di cui 92 famiglie, 758 generi e 2.912 specie (IPK). Le responsabilità della banca del germoplasma sono di collezionare, conservare e caratterizzare le risorse genetiche delle piante in loro possesso, condurre ricerche e sviluppare metodi di conservazione efficienti (IPK). A tal proposito, tra le attività svolte, la banca del germoplasma IPK ha sviluppato dei duplicati di sicurezza di circa 54.000 accessioni conservate (≅ 36%), successivamente depositati alla "Global Seed Vault" a Svalbard, Norvegia (IPK). Le risorse genetiche conservate sono spesso corredate da informazioni fondamentali e metadati, come il passaporto (e.g. stato tassonomico, sito di raccolta, provenienza del materiale; Milner et al. 2019) ed eventualmente, ma non sempre, dati agronomici, caratteristiche biochimiche ed informazioni riguardanti la sequenza del DNA ed il genotipo (e.g., whole genome sequencing data; WGS, varianti molecolari da esperimenti di genotyping by sequencing; GBS) (https://www.ipk-gatersleben.de/en/). L'identificazione di duplicati e la gestione di accessioni ridondanti all'interno e tra le raccolte di germoplasma è una delle sfide principali che devono affrontare le banche del germoplasma (Hintum et al. 1995).

I dati molecolari possono essere impiegati per investigare la somiglianza tra accessioni depositate entro una banca o in banche del germoplasma differenti, e identificare potenziali duplicati (Milner et al. 2019).

Ciò si riflette nella possibilità di utilizzare dati molecolari per una più efficiente conservazione delle risorse genetiche, oltre alla possibilità di integrare ed eventualmente correggere dati fenotipici e di passaporto (Milner et al., 2019). L'analisi del genoma mediante marcatori molecolari o l'intera sequenza permette quindi di caratterizzare le risorse genetiche e produrre un passaporto molecolare (Milner et al. 2019).

Figura 5: Conservazione di accessioni cerealicole nella banca del Germoplasma presso IPK.

Tra gli aspetti tecnici da considerare nei processi di conservazione, la percentuale di umidità dei semi risultante particolarmente rilevante; possiamo distinguere semi ortodossi, conservati ad un basso livello di umidità e temperatura senza perdere vitalità, dai semi recalcitranti, i quali perdono vitalità a percentuali di umidità inferiori al 12-13% (Biodiversity International 2009). Prima che il seme venga confezionato e stoccato, occorre determinarne la vitalità attraverso una prova di germinabilità e successivamente l'assenza di patogeni e/o malattie (Biodiversity International 2009) (Rao et al. 1994).

La conservazione delle accessioni può avvenire attraverso l'istituzione di "*Base Collection*" e "*Active Collection*" (Salgotra et al. 2023). Nel caso di una *Base Collection*, i campioni di seme sono stoccati per il loro massimo tempo di vita a temperature che vanno dai -18 a -20 C° (Jose et al. 2018), con una umidità tra il 3% ed il 7%, dipendentemente dalla specie (Salgotra et al. 2023). Per quanto riguarda invece una *Active Collection* i semi sono stoccati per un uso immediato con un periodo massimo di 20 anni, vitalità almeno del 65% e un'umidità tra il 7% e 11% per i semi con una buona conservabilità, mentre per quelli con una

bassa conservabilità l'umidità si mantiene tra 3% e 8% (Salgotra et al. 2023). In linea generale, i campioni di seme delle accessioni possono andare incontro a strategie di conservazione nel breve, medio e lungo periodo, con differenti temperature ed umidità di gestione (Salgotra et al. 2023). Per una efficiente conservazione delle risorse genetiche, le banche del germoplasma possono definire delle "*Core Collection*" ovvero delle collezioni costituite da sottocampioni ottenuti dalle "*Base Collection*". L'obiettivo delle *Core Collection* è quello di costituire delle collezioni in cui si massimizzi la variabilità genetica e fenotipica presente nelle collezioni originali, col supporto di dati di passaporto, e limitando quindi la ridondanza genetica (Gu et al. 2023). L'utilizzo di *Core Collection* permette quindi di caratterizzare un set di materiali relativamente contenuto (e.g., 10% della collezione di base), conservando al tempo stesso la variabilità presente delle collezioni originali; ciò permette di facilitare la caratterizzazione, l'utilizzo e la conservazione delle risorse genetiche stoccate nelle banche del germoplasma (Gu et al. 2023, Brown 1989).

Il DNA contenuto all'interno del seme può andare incontro a degradazione con il passare del tempo, oltre ad una generale diminuzione della germinabilità della semente; si rende quindi necessaria una riproduzione periodica del materiale stoccato prima che questo perda la vitalità, così da poter sostituire il vecchio seme con il nuovo seme vitale (Rajasekharan, 2015). Il seme rappresenta uno dei materiali maggiormente conservati nelle banche del germoplasma, ma ad oggi con i progressi in campo biologico, biotecnologico e genomico possiamo individuare, tra il materiale che può essere conservato come risorsa di variabilità allelica, anche polline, DNA, librerie di DNA genomico, RNA, e tessuti vegetali (Figura 4), mediante specifiche condizioni di stoccaggio (Rajasekharan. 2015, Salgotra et al. 2023, Pandotra et al. 2015).

Ai fini della conservazione di organi e tessuti è stata implementata la tecnica della crioconservazione (Figura 4), la quale consiste nello stoccaggio di espianti vegetali provenienti da coltura in vitro (gemme, meristemi, embrioni, colture cellulari) in azoto liquido (-196°C) (CNR).

1.2 **Valorizzazione delle risorse genetiche nelle leguminose alimentari**

1.2.1 *Caratterizzazione delle risorse genetiche*

Il materiale genetico conservato nelle banche del germoplasma rappresenta una fonte inestimabile di variabilità e diversità genetica. Tuttavia, data l'enorme quantità di materiale stoccato, l'identificazione di opportune strategie di caratterizzazione, come la costituzione di Core Collection, diventa fondamentale. La caratterizzazione delle risorse genetiche a sua volta fornisce informazioni cruciali sulla loro diversità, in modo da definire opportune strategie di

conservazione e per un loro eventuale impiego in agricoltura ed in programmi di miglioramento genetico. Le risorse genetiche possono essere caratterizzate su due livelli: caratterizzazione fenotipica e genotipica. Con riferimento alla caratterizzazione delle risorse genetiche, si deve tener conto del fatto che nelle banche del germoplasma la singola accessione è in realtà spesso costituita da materiali eterogenei. Un aspetto rilevante quando si conducono studi sul fenotipo e sul genotipo, e soprattutto studi di correlazione tra marcatori molecolari e caratteri fenotipici sulle popolazioni (si faccia riferimento al prossimo paragrafo; identificazione di QTL e geni di interesse), è la costituzione di seme derivato per discendenza da singolo seme (SSD), in cui l'accessione derivata è rappresentata da un unico genotipo. Questo approccio, basato sull'autofecondazione permette di ottenere a partire da un seme dell'accessione di partenza, ed in particolar modo in specie prevalentemente autogame, genotipi altamente omozigoti (i.e., linee pure) su cui si possono fare inferenze nella correlazione tra genotipo e fenotipo (Bellucci et al., 2023). In particolar modo, una comune strategia è quella di assegnare dei DOI (i.e. Digital Object Identifier) alle accessioni, in modo che queste possano essere identificate e documentate in maniera univoca e permanente, facilitando anche lo scambio delle risorse genetiche tra i portatori d'interesse e la loro tracciabilità.

1.2.1.1 Caratterizzazione Fenotipica

La fenotipizzazione consiste nell'applicazione di metodologie e protocolli per misurare e descrivere la manifestazione di un tratto specifico a livello di singolo individuo e a livello di popolazione per investigare la variabilità fenotipica disponibile; caratteristiche fenotipiche possono riguardare caratteri morfologici (e.g., habitus di crescita, forma di un organo come la foglia, dimensione dei semi), fenologici (e.g., precocità di fioritura e maturazione), ed agronomici (e.g., produttività). In linea generale, la fenotipizzazione può riguardare specifici organi e su ampie popolazioni interessare diversi stadi di sviluppo della pianta (Singh et al. 2016). La caratterizzazione fenotipica può riguardare inoltre risposte a stress biotici ed abiotici, architettura delle radici e della chioma, fisiologia, aspetti ecologici (e.g., interazione tra piante e piante con microorganismi), caratteristiche qualitative delle parti edibili, efficienza fotosintetica etc… In passato questi caratteri erano raccolti manualmente; oggi, invece, con l'ausilio di computer, tecnologie avanzate ed intelligenza artificiale, sono disponibili approcci e strategie per la "fenotipizzazione ad alto rendimento" (i.e., high throughput phenotyping) (Li et al. 2014). Questa nuova frontiera si avvantaggia dell'utilizzo della robotica (Xu et al. 2022), del controllo avanzato dei fattori ambientali (Heuermann et al. 2023) e della diagnosi mediante immagine per determinare tratti relativi alla pianta (Li et al. 2014), come ad esempio

l'architettura radicale (Trachsel et al. 2011; Zhu et al. 2011). La tecnologia più datata per la fenotipizzazione mediante immagine è quella che fa riferimento al campo delle radiazioni visibili (≅400-700 nm) (Figura 6) ed è stata molto utilizzata per comprendere la risposta allo stress idrico, nonché per la caratterizzazione di aspetti dinamici come la morfologia, l'architettura ed il tasso di crescita (Mishra et al. 2016). Un'evoluzione di questo approccio ha riguardato l'integrazione delle immagini termiche catturate nello spettro dell'infrarosso (≅750-1300 nm) (Mishra et al. 2016) (Figura 6).

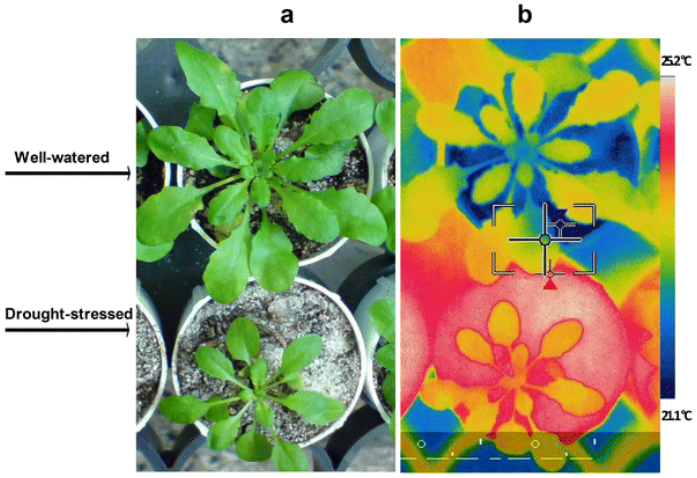


Figura 6: L'utilizzo di immagini termiche nello spettro dell'infrarosso per studi di caratterizzazione fenotipica high throughput. a) immagine digitale RGB; b) immagine termica a infrarossi. Le immagini mettono a confronto due individui di Arabidopsis thaliana in diverse condizioni idriche (A e B, ben irrigate, in alto; stress idrico, in basso). Figura tratta da Mishra et al. (2016)

Attraverso l'utilizzo della spettroscopia ad immagine è possibile captare il segnale di riflettanza della pianta e caratterizzare l'attività fotosintetica in termini di capacità di assorbimento dell'energia luminosa dei pigmenti (Mishra et al. 2016).

La fenotipizzazione non si limita solo a caratteri visibili, ma include lo studio di caratteristiche istologiche, fino all'analisi del fenotipo molecolare (i.e., metabolomica e trascrittomica). L'impiego delle tecnologie -*omiche* ha permesso di raggiungere livelli di caratterizzazione particolarmente dettagliati, e tra queste, troviamo la Genomica (di cui si parlerà nel dettaglio nel prossimo paragrafo), la Trascrittomica, la Proteomica, e la Metabolomica. Precisamente la trascrittomica si occupa dello studio del trascrittoma, ovvero l'insieme di tutti gli RNA messaggeri (mRNA), investigando quindi il grado di espressione genica (Dong et al. 2013; Yang et al. 2014). La proteomica concerne lo studio quali-quantitativo del proteoma, ossia la totalità delle proteine espresse da una cellula definendone

la struttura e la funzione (Collotta 2018). La metabolomica si occupa della caratterizzazione di struttura, funzione e ruolo dei metaboliti cellulari (Patti GJ et al. 2012).

1.2.1.2 Caratterizzazione genotipica

La caratterizzazione genotipica (genotipizzazione) consiste nello studio dei polimorfismi del DNA; analizzando le differenze genetiche tra individui e popolazioni, è possibile condurre studi di genetica di popolazione, genomica di popolazione, ed utilizzare le varianti molecolari, in associazione con dati fenotipici, per condurre studi di associazione genotipo-fenotipo (Genome Wide Association Studies; GWAS) (Scheben et al. 2017). Nel campo della genetica agraria la genotipizzazione gioca un ruolo cruciale sia nella identificazione di QTL per caratteri di interesse, sia nello sviluppo di metodologie di miglioramento genetico assistito da marcatori molecolari (Torkamaneh et al. 2018). Un marcatore molecolare è una sequenza di DNA la cui posizione sul cromosoma è conosciuta, ed è associato con un particolare gene o tratto (Al-Samarai et al. 2015). I marcatori molecolari si basano sulle differenze tra individui nella sequenza del DNA genomico (polimorfismi dovuti a mutazioni) e la valutazione di tali differenze ci fornisce un sistema di analisi di alta precisione (Lorenzetti et al. 2021)

I marcatori molecolari oggi più utilizzati per la caratterizzazione molecolare della diversità genetica presente entro e tra le popolazioni sono i ***DArT*** *(Diversity Array Technology,* ***SSRs*** *(Simple Sequence Repeats) e* ***SNPs*** *(Single Nucleotide Polymorphisms)* (Salgotra et al. 2023).

I dati ottenuti dall'analisi dell'intero genoma o da singoli marcatori molecolari, come riportato nei precedenti paragrafi, permettono l'identificazione di duplicati ed errori nei dati di passaporto (mediante la costituzione di un "Passaporto Molecolare"), nelle collezioni di risorse genetiche; queste informazioni possono supportare lo sviluppo di "Core Collection" per studi più approfonditi (Milner et al. 2019), e facilitare la conservazione e la caratterizzazione ulteriore di tali risorse.

Un importante passo in avanti verso una migliore correlazione tra genotipo e fenotipo è stato reso possibile grazie all'avvento delle tecniche di nuova generazione per il sequenziamento del genoma (Verma et al. 2017). Infatti, si distinguono tre generazioni per le tecniche di sequenziamento (Verma et al. 2017). Il sequenziamento di prima generazione, comprende il metodo Sanger (sequenziamento per sintesi) ed il metodo Maxam-Gilbert (sequenziamento per scissione) (Verma et al. 2017). L'evoluzione di tali metodi ha dato poi vita alle tecniche di "Next Generation Sequencing" (NGS) che caratterizzano sia la seconda che la terza generazione delle tecniche di sequenziamento (Verma et al. 2017). Le tecniche NGS presentano importanti vantaggi rispetto al metodo Sanger in termini di costo e velocità

di sequenziamento (Verma et al. 2017), permettendo l'acquisizione di dati per l'intero genoma su ampi set e collezioni di risorse genetiche.

Come sopra riportato, i dati molecolari combinati a dati disponibili sul fenotipo permettono di investigare la diversità e identificare geni d'interesse impiegabili in programmi di miglioramento genetico per caratteri quali-quantitativi (Milner et al. 2019).

Con riferimento al miglioramento genetico e alle sfide future dell'agricoltura, tra tutte le leguminose alimentari possono fornire un contributo fondamentale nel progresso dell'agricoltura e dell'industria alimentare verso un futuro sostenibile (Magrini et al. 2018). Nel 2019, il report dell' IPCC intitolato "Climate Change and Land" ha indicato che la transizione verso una dieta basata su vegetali potrebbe rappresentare una grande opportunità per l'adattamento e la mitigazione del cambiamento climatico e allo stesso tempo creare benefici per quello che riguarda la salute umana (Bellucci et al. 2021).

1.2.2 *Leguminose e importanza nei sistemi agroalimentari moderni*

Le leguminose o fabacee sono una famiglia di piante dicotiledoni dell'ordine delle Fabales. Come riportato nell'elaborato di tesi triennale (Paolinelli. 2022), le leguminose presentano fiori ciclici, eteroclamidi, pentameri, monoclini, (raramente diclini), actinomorfi o zigomorfi (Cortesi, 1933). Le leguminose vengono così definite in quanto producono frutti detti baccelli o legumi, di forma allungata e costituiti da due valve entro cui sono allineati i semi. Le leguminose possono comprendere specie erbacee annue o perenni, arbusti o alberi con foglie per lo più composte (Cortesi, 1933). La famiglia delle Fabacee è una delle più grandi famiglie delle piante vascolari, con circa 12.000 specie riunite in 430 generi (Cortesi, 1933). Si dividono in tre sottofamiglie: *Mimosoideae*, *Caesalpinioideae* e *Faboideae* (*Papilionoideae*) (Stevens, 2017; Cortesi, 1933). Le Faboideae comprendono la maggior parte delle specie di importanza alimentare, soprattutto per il grande apporto proteico dei loro semi, come la soia (*Glycine max*), il pisello (*Pisum sativum*), il cece (*C. arietinum*), il fagiolo comune (*P. vulgaris*), la lenticchia (*L. culinaris*), il lupino (*L. albus e L. mutabilis*) e l'arachide (*Arachis hypogea*).

Le leguminose sono considerate in agronomia come delle colture miglioratrici del terreno in virtù delle loro capacità; infatti, l'apparato radicale è capace di instaurare un rapporto simbiotico con i batteri azotofissatori (e.g. *Rhizobium*, *Sinorhizobium*, *Azorhizobium*, *Bradyrhizobium*, *Mesorhizobium*; Patriarca et al. 2002) i quali operano la riduzione dell'azoto atmosferico (N2) in azoto ammoniacale (NH3). Parte dell'azoto ammoniacale sintetizzato nel processo di azoto fissazione è reso disponibile per il metabolismo della pianta mentre parte

raggiunge il terreno diventando disponibile per le altre colture presenti in campo ma anche per quelle che succedono in rotazione la coltura leguminosa.

In virtù di queste capacità e dell'effetto che hanno sulla struttura, sulla fertilità chimica e biologica del terreno le leguminose sono utilizzate in agricoltura in rotazioni colturali prima e/o dopo colture depauperanti come i cereali, rappresentando un vantaggio per la produttività e sostenibilità dell'agro-ecosistema (Bellucci et al. 2021).

Per quanto riguarda invece l'aspetto nutrizionale, i legumi sono un alimento prezioso sotto perché presentano un alto livello proteico (20-45%), con la presenza di alcuni aminoacidi essenziali, e contengono carboidrati complessi (30-60%) e fibre (5-37%) (Maphosa et al., 2017). I legumi hanno un basso livello di grassi, ad eccezione delle arachidi, contengono minerali essenziali (ferro, calcio, zinco),vitamine (e.g., vitamine del gruppo B) (Maphosa et al., 2017) e composti bioattivi con azione antiossidante (Carbonaro et al 2015; Maphosa e t al. 2017).

1.3 **Approcci partecipativi nella conservazione decentralizzata delle risorse genetiche**

Gli approcci partecipativi prevedono una serie di tecniche e attività, volte a condividere e coinvolgere senza distinzioni i membri della comunità nei processi di ricerca fornendo un'opportunità unica per innovare il paradigma alla base della produzione di sapere scientifico (Roque et al. 2022). Diversi sono gli approcci partecipativi e tra questi, vi sono gli esperimenti di scienza dei cittadini (Citizen Science Experiment), termine coniato dal sociologo scientifico Alan Irwin e l'ornitologo Richard Bonney (Irwin. 1995; Bonney et al. 1996). Irwin (1995) pubblica il libro "Citizen Science: A Study Of People, Expertise and Sustainable Development" dove definisce la "scienza del cittadino" come una scienza che serve gli interessi del cittadino performata dallo stesso ("Science for the People" e "Science by the People"). Richard Bonney et al. (2016) definisce il "Citizen Science" come un progetto scientifico dove il cittadino fornisce i dati raccolti agli scienziati ed in cambio acquisisce capacità scientifiche. Bonney et al. (2016) riassume questo processo come "A two-way street" indicando lo scambio tra scienziati e cittadini; i primi ricevono una quantità elevata di dati grazie al lavoro di caratterizzazione dei cittadini, altrimenti non ottenibile dal singolo scienziato; i secondi acquisiscono una maggiore consapevolezza e conoscenza delle risorse studiate. Il termine "Citizen Science" è utilizzato per indicare un ampio spettro di pratiche e finalità: dai cittadini che contribuiscono con l'elaborazione di grandi dataset mediante i propri computer (e.g., SETI@Home), ai naturalisti amatoriali che raccolgono dati sull'osservazione di uccelli (i.e., eBird), cittadini che mappano gli inquinanti nelle città (e.g., City Sense),

persone che classificano immagini on-line delle galassie da casa (e.g., Galaxy Zoo), pazienti che condividono osservazioni, sintomi ed esperienze sulla loro salute (e.g., PatientsLikeMe) e i biohacker che cercano di produrre insulina nei laboratori comunitari (e.g., Counter Culture Labs) (Strasser et al. 2018). L'avanzare degli approcci partecipativi nel campo della ricerca rappresenta una sfida, non solo per la scienza stessa ma anche per il corrente ordine sociale, offrendo un esempio di co-produzione di informazioni scientifiche e nuovi rapporti sociali (Shapin and Schaffer. 1985; Jasanoff. 2004). Tutto questo punta ad una potenziale trasformazione nella interazione tra scienza e cittadino "comune" (Strasser et al. 2018) dove si sfida il moderno regime di produzione di conoscenza scientifica basato sulla netta separazione tra scienziato professionale e cittadino (Strasser et al. 2018).

L'ambiziosità del termine "Citizen Science" si presenta con tre importanti promesse: "Democratizzare la scienza" (distribuzione del potere tra tutti i cittadini, anche quelli meno rappresentati) (Strasser et al. 2018), "educare i cittadini alla scienza" (rendere i cittadini più consapevoli dei processi di ricerca e sviluppo, riducendo il timore verso l'innovazione) (Strasser et al. 2018) e produrre nuova "scienza" (generare un massivo contingente di dati grazie all'ampia inclusione del pubblico) (Strasser et al. 2018).

I singoli programmi di Citizen Science richiedono supervisione, coordinazione, sviluppo e affinamento dei protocolli, allenamento, un'infrastruttura per la gestione dei dati e supporto finanziario (Bonney et al. 2009; Cohn 2008). Per un efficiente raggiungimento e condivisione dei risultati risulta fondamentale non solo la singola organizzazione di un progetto ma l'intera cooperazione e collaborazione tra i diversi progetti di Citizen Science che trattano obiettivi simili (Newman et al. 2011).

Questi approcci partecipativi possono essere ampiamente applicati al mondo dell'agricoltura per affrontare quelle che sono le costanti e crescenti pressioni che spingono l'intero mondo agricolo verso un veloce cambiamento della sua architettura (Van Etten et al. 2019). La necessità di innovazione in campo agricolo non solo è riferita allo sviluppo di nuove tecnologie da applicare in campo di robotica e/o di agricoltura di precisione, bensì con le tendenze socioeconomiche odierne come l'aumento demografico, urbanizzazione, cambiamenti dietetici, cambiamento climatico e quindi la necessità di una produzione sostenibile, diventa fondamentale definire e creare nuove varietà adatte a questi nuovi contesti. Tradizionalmente il processo è gestito nella sua completezza da ricercatori e genetisti professionisti in ambito privato e/o pubblico e richiede diverse tempistiche a seconda delle metodologie di miglioramento genetico impiegate. Nella maggior parte dei casi queste tempistiche risultano lunghe e quindi problematiche al fine del superamento di criticità

climatiche, in considerazione della difficoltà di poter caratterizzare ampi set di risorse genetiche conservate nelle banche del germoplasma. Il processo deve essere quindi velocizzato e la varietà prodotta deve incontrare più che mai le esigenze degli agricoltori e dei territori, sia quelli adattati ad agricolture intensive sia verso gli areali marginali di più difficile coltivazione (i.e. selezione decentralizzata. Simmonds 1991). infatti, la ricerca e lo sviluppo in campo agricolo, ha portato a dei sensibili aumenti di produzione ed approvvigionamento alimentare in molte aree ad eccezione delle aree marginali in paesi in via di sviluppo con alti livelli di povertà (Ceccarelli. 2006). Questo sbilanciamento è dovuto al fatto che la maggior parte delle fasi di selezione e sviluppo di nuove varietà avvengono in un ristretto numero di stazioni sperimentali (variabilità ambientale poco rappresentata) e le decisioni sono prese unilateralmente dai miglioratori senza un approccio integrato (Ceccarelli et al. 2007), portando all' esclusione di materiale di breeding potenzialmente utile in aree marginali poco rappresentate. Questa situazione rende quindi indispensabile un elevato numero di ambienti rappresentati, in modo da rendere la selezione più efficiente a livello sito-specifico ed in linea con la situazione agronomica, tecnologica e socioeconomica di un dato territorio (Ceccarelli et al. 2007). Per rispondere a queste esigenze, oltre agli esperimenti di Scienza dei Cittadini (si veda nel dettaglio il prossimo paragrafo), ben si presta l'approccio partecipativo nel miglioramento genetico (Participatory plant breeding; PPB) il quale consiste nella collaborazione tra il miglioratore e l'utilizzatore del prodotto finito (nuova varietà) (Ceccarelli et al. 2007). Questo tipo di collaborazione può essere distinta in cinque tipologie a seconda del grado di partecipazione degli agricoltori come riassunto nella successiva tabella 2 (Ashby. 2009).

Tabella 2: Strategie e tipologie di approcci partecipativi nella conduzione di esperimenti scientifici. (Ashby. 2009)

Convenzionale	Consultativo	Collaborativo	Collegiale	Sperimentazione degli agricoltori
Lo scienziato prende decisioni autonomamente senza la partecipazione di agricoltori.	Gli scienziati prendono decisioni autonome ma in comunicazione con gli agricoltori.	Scienziati ed agricoltori hanno la capacità di prendere decisioni riguardo al processo di miglioramento genetico.	Gli agricoltori prendono decisioni riguardo al processo di breeding collegialmente tra di loro in comunicazione con gli scienziati.	Non si ha la partecipazione di scienziati, gli agricoltori fanno le loro scelte individualmente o collegialmente tra di loro.

Un esempio pratico di scienza del cittadino per indagare l'adattabilità varietale all'ambiente è stato quello condotto nel 2016 dall'Università di Hohenheim e Taifun-Tofu GmbH, i quali

realizzarono il progetto di “Citizen Science” “1000 Gardens -The Soybean Experiment” (Würschum et al. 2019). I cittadini in questo caso si impegnavano a coltivare e valutare delle “Breeding Lines” di un programma di miglioramento genetico già avviato che riguardava la soia (Würschum et al. 2019). Gli obiettivi del progetto erano di collezionare dati per analisi genetiche per determinare l’adattamento delle diverse linee di soia all’ambiente e clima della Germania e migliorare l’immagine pubblica dei legumi (soia specialmente) sottolineando i loro vantaggi come coltura centrale per una transizione verso un regime produttivo agricolo ed alimentare sostenibile (Würschum et al. 2019).

L’adozione del Citizen science ha permesso di valutare le linee di soia in un numero elevato di ambienti diversi tra loro permettendo l’identificazione di genotipi diversamente adattabili agli areali tedeschi (Würschum et al. 2019). È stato riscontrato che le principali motivazioni dietro la partecipazione dei cittadini erano che trovavano l’esperimento interessante, traevano interesse dall’imparare nuove cose ed essere parte attiva contribuendo alla ricerca scientifica (Würschum et al. 2019). Grazie all’esperimento condotto da Würschum et al. (2019) si è potuto affermare che i principali fattori che influenzano la partecipazione dei cittadini sono la concezione delle problematiche ambientali e la facilità/difficoltà con la quale queste vengono presentate al pubblico (Würschum et al. 2019). Nel prossimo paragrafo si discuterà nel dettaglio del progetto INCREASE, per la caratterizzazione di risorse genetiche di leguminose alimentari, il quale include, tra le varie analisi ed approcci, un interessante esempio di Esperimento di Scienza dei Cittadini come strategia per la caratterizzazione e la conservazione decentralizzata delle risorse genetiche.

1.3.1 *INCREASE*

Figura 7: Logo del progetto INCREASE (https://www.pulsesincrease.eu/)

Nell’ambito della gestione, caratterizzazione e valorizzazione delle risorse genetiche, è nato nel 2020 il progetto INCREASE (figura 7) (Intelligent Collections of Food Legumes Genetic Resources for European Agrofood Systems) (https://www.pulsesincrease.eu/) all’interno del programma europeo Horizon2020 (Sito INCREASE). Il progetto nasce con l’obiettivo di migliorare e valorizzare la gestione e l’utilizzo delle risorse genetiche delle leguminose alimentari, cruciali per la sostenibilità, sicurezza alimentare e salute umana (sito INCREASE), implementando in questo senso anche strategie di conservazione decentralizzata.

1.3.2 *Specie leguminose di interesse alimentare studiate nel progetto INCREASE*

Il progetto INCREASE è incentrato su quattro specie leguminose, rappresentate da Fagiolo Comune (*Phaseolus vulgaris*), Cece (*Cicer arietinum*), Lenticchia (*Lens culinaris*), Lupino (*Lupinus albus* e *Lupinus mutabilis*). La scelta delle leguminose e nello specifico di queste quattro specie sta nel loro valore potenziale per la produzione sostenibile di cibo strettamente legato alla tradizione e la cultura europea (Bellucci et al. 2021).

1.3.2.1 Fagiolo

Figura 8: Immagini rappresentative dello sviluppo e delle caratteristiche fenotipiche nella specie Phaseolus vulgaris. A) plantula appena emersa di fagiolo dove possiamo vedere i due cotiledoni; B) due baccelli di fagiolo; C) fiori di fagiolo; D) pianta di fagiolo a crescita determinata (nano); E) pianta di fagiolo a crescita indeterminata (rampicante).

Il fagiolo comune (*Phaseolus vulgaris*) (Figura 8) è una specie erbacea diploide (2n=2x=22; dimensioni del genoma ≅520Mbp), annuale e prevalentemente autogama ed è la leguminosa da granella più importante per il consumo umano nel mondo. In Europa, negli ultimi anni si è potuto notare un sensibile incremento dell'interesse pubblico verso le

leguminose ed in particolare verso il fagiolo; infatti, oggi si possono contare circa 544.330 ha coltivati a fagiolo comune con una produzione di 1.9mt (INCREASE). I fagioli sono largamente coltivati per la produzione di granella (Dry Bean) o per la produzione del baccello immaturo alla raccolta e consumato fresco (Green Bean; fagiolino verde). *Phaseolus vulgaris* è parte del genere *Phaseolus* insieme ad altre 70 specie. All'interno del genere *Phaseolus* si possono distinguere otto cladi principali per via della diversa morfologia, ecologia e distribuzione bio-geografica (INCREASE). *P. vulgaris* ha avuto origine in Mesoamerica (Bitocchi et al. 2012; Ariani et al. 2018; Desiderio et al. 2013; Rendón-Anaya et al. 2017; Schmutz et al. 2014) circa 4 milioni di anni fa (INCREASE) e successivamente, attraverso fenomeni migratori si è distribuito nelle zone montuose dell'America latina e nel nord-ovest dell'Argentina (Toro et al. 1990).

Il fagiolo comune è caratterizzato da tre pool genici selvatici con diversa distribuzione eco-geografica: Mesoamerica e Ande che sono i due maggiori pool genici ed includono oltre alle forme selvatiche, genotipi domesticati; ed il pool genico "Nord Perù-Equador" caratterizzato da una ristretta area di distribuzione (Freyre et al. 1996; Kami et al. 1995; Gepts et al. 1999; Cortinovis et al. 2021). Il Fagiolo comune, inoltre, è stato soggetto a due eventi di domesticazione paralleli ed indipendenti rispettivamente in Mesoamerica e nelle Ande (Bitocchi et al. 2012), che hanno contribuito alla riduzione della diversità a livello nucleotidico; inoltre, si è osservata una riduzione della diversità di espressione genica associata alla domesticazione nel pool Mesoamericano, come dimostrato da Bellucci et al. (2014). Allo stesso tempo è stato possibile osservare un aumento della diversità nucleotidica a livello di alcuni loci responsabili dell'adattamento ambientale a fattori biotici ed abiotici (Bellucci et al. 2014; Bitocchi et al. 2017). Il fenomeno di domesticazione primaria avvenuto in Mesoamerica e nelle Ande ha contribuito all'ottenimento di diverse razze di fagiolo domesticate, così come descritto da Singh et al. (1991), sulla base di dati di passaporto ed evidenze fenotipiche legate alla morfologia e fisiologia delle piante e del seme (e.g., dimensione del e pattern di colorazione del seme, habitus di crescita, sensibilità al fotoperiodo, colore del fiore etc.), si possono identificare tre razze Mesoamericane (Durango, Jalisco e Mesoamerica) e tre razze Andine (Perù, Nueva Granada e Chile). In un recente lavoro (Bellucci et al., 2023), mediante l'utilizzo di una analisi ADMIXTURE (Alexander et al. 2009; Figura 9) su dati WGS di accessioni Americane ed Europee di fagiolo, è stata osservata con una classificazione K=2 una principale separazione delle accessioni in due gruppi genetici, che sulla base dei dati di passaporto, corrispondono ad accessioni Mesoamericane (Cluster 2 in Figura 9A) ed Andine (Cluster 1 in Figura 9A); inoltre, quando gli autori hanno investigato

la sottostruttura genetica delle accessioni Americane, ADMIXTURE ha permesso di identificare 2 gruppi genetici entro il pool genico Mesoamericano (M1 e M2) e 3 gruppi genetici entro il pool Andino (A1, A2, e A3) (Figura 9B). Infine, Bellucci et al. (2023) identificano una chiara corrispondenza tra le razze descritte da Singh et al. (1991) su base fenotipica e i 5 gruppi genetici così come identificati su base molecolare. Si è stabilita la seguente corrispondenza: il gruppo/pool genico M1 comprende accessioni delle razze Durango e Jalisco, il pool genico M2 corrisponde ad accessioni della razza Mesoamerica, il pool genico A1 corrisponde ad accessioni della razza Nueva Granada, il pool genico A2 corrisponde ad accessioni della razza Perù, ed il pool genico A3 corrisponde ad accessioni della razza Chile.

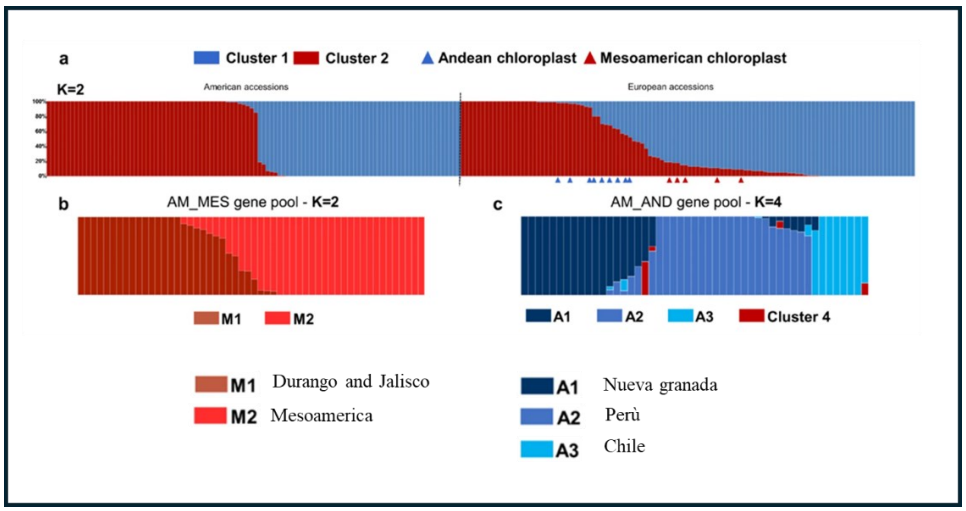


Figura 9: Risultati dell'analisi ADMIXTURE condotta da Bellucci et al. (2023) su un set di accessioni Americane ed Europee di fagiolo. a) Analisi ADMIXTURE che mostra il livello di ibridazione nel genoma in accessioni di fagiolo (americane a SX e europee a DX; dati di passaporto) sulla base di due cluster principali identificati su base molecolare mediante WGS (cluster 1, corrispondente a materiali di provenienza Andina e cluster2 corrispondente a materiali di provenienza Mesoamericana) da Bellucci et al. 2023. A sinistra possiamo notare come i due pool genici siano rimasti per lo più separati con solo alcune accessioni interessate da processi di ibridazione ed introgressione; nelle accessioni europee, dove la rottura di barriere geografiche ha permesso l'ibridazione dei due pool genici, si osserva un alto numero di accessioni cosiddette "admixed". b) "ADMIXTURE" analysis che mostra il livello di ibridazione nel genoma in accessioni di fagiolo comune mesoamericano sulla base delle due razze identificate su base molecolare mediante WGS da Bellucci et al (2023) (M1 e M2). C) "ADMIXTURE" analysis che mostra il livello di

ibridazione nel genoma in accessioni di fagiolo comune andino sulla base delle due razze identificate su base molecolare mediante WGS da Bellucci et al (2023) (A1, A2, A3 e Cluster4).

Sulla base delle evidenze molecolari e fenotipiche, Bellucci et al. (2023) hanno suggerito un modello per la domesticazione del fagiolo e la successiva introduzione in Europa di genotipi di questa specie (Figura 10). In particolare, alcuni individui della razza M1, che è stata la prima ad essere domesticata dai selvatici hanno dato origine alla razza M2 (Mesoamerica) situata più a sud rispetto a M1; A2, domesticata a partire dai selvatici, ha probabilmente dato origine ad individui della razza A1 (Nueva Granada) e A3 (Chile) (Bellucci et al. 2023) (Figura 9A). Questo modello è supportato dall'ipotesi che le prime razze siano state parzialmente domesticate, rimanendo però sensibili al fotoperiodo (Figura 10). La riduzione o perdita della sensibilità al fotoperiodo rappresenterebbe un passaggio fondamentale per la successiva introduzione ed efficace disseminazione di alcune razze genetiche in Europa. In particolare, la razza Nueva Granada (A1) potrebbe essere stata la prima razza introdotta e diffusa per l'assenza di sensibilità al fotoperiodo (Bellucci et al., 2023, Figura 10A e 10B). Inoltre, sulla base dei dati molecolari, non si riscontra la diffusione del gruppo genetico A2 (razza Perù) in quanto sensibile al fotoperiodo (Figura 10B), mentre Bellucci et al. (2023) ipotizzano che l'elevato livello di introgressione e ricombinazione osservato negli individui europei prevalentemente assegnati al gruppo genetico M1, sia uno dei meccanismi alla base dell'introduzione di questo gruppo genetico in Europa (Figura 10).

La domesticazione secondaria è stato un evento chiave per l'adattamento del fagiolo comune in Europa, ed è stata quindi associata alla riduzione o perdita della sensibilità al fotoperiodo; infatti, grazie ai viaggi di Cristoforo Colombo a partire dal 1492 il fagiolo comune ha raggiunto il vecchio mondo (Figura 10A e 10B). Come riportato, la conseguente rottura delle barriere spaziali geografiche tra il pool mesoamericano e andino in Europa, per la razza M1 è stato possibile un processo di adattamento. L'elevato livello di ibridazione tra gene pool Andino a Mesoamericano, realizzatosi in centri secondari di domesticazione come l'Europa è evidente dai dati di ADMIXTURE di Bellucci et al. (2023) (Figura 9a; European accessions), ma era già stato riportato anche da altri autori (Cortinovis et al. 2021; Bellucci et al. 2023; Papa 2007; Gepts et al. 1988; Santalla et al 2002; Ariani et al. 2018).

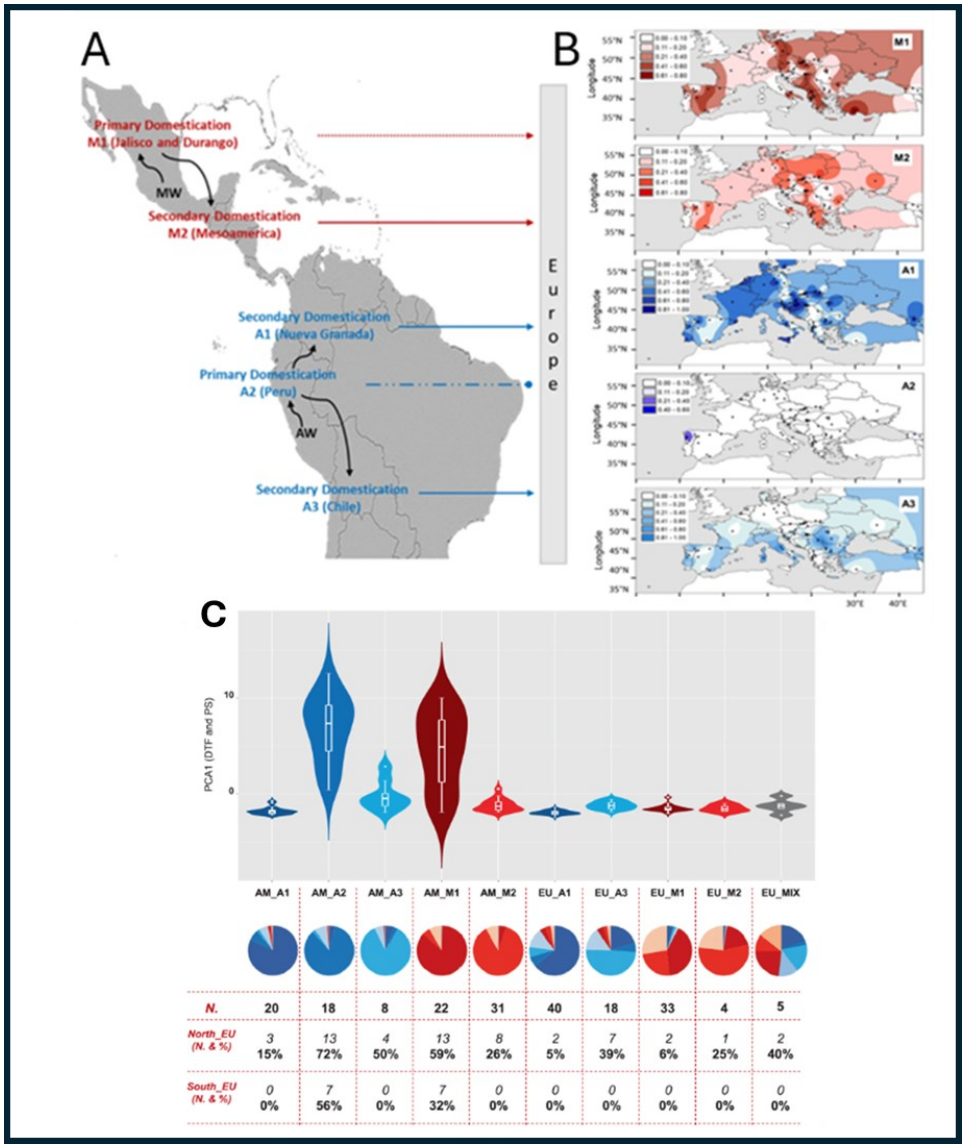


Figura 10: La perdita della sensibilità al fotoperiodo e meccanismi di introgressione e ricombinazione alla base dell'adattamento del fagiolo comune in seguito alla sua introduzione in Europa. A) Eventi di domesticazione primaria e secondaria con i relativi fenomeni migratori nei pool genici mesoamericano e andino di fagiolo comune. La perdita completa o parziale di sensibilità al fotoperiodo durante la domesticazione secondaria è stato un fattore rilevante per l'introduzione delle razze A1 (Nueva Granada), A3 (Chile) e M2 (Mesoamerica) in Europa (con la freccia intera si indica la loro introduzione e diffusione in Europa); Freccia tratteggiata per il pool genico M1 indica l'introduzione e diffusione del materiale genetico grazie a fenomeni di introgressione del genoma di altre razze; linea tratteggiata senza punta per il pool genico A2 indica mancata introduzione e/o diffusione della razza A2 sensibile al fotoperiodo B) Interpolazione spaziale della distribuzione geografica delle "ancestry" su base molecolare, che mostra la distribuzione

delle razze M1, M2, A1, A2, A3 in Europa. Possiamo notare come il pool genico A2 (razza Perù) non sia stata efficientemente introdotta in Europa per via della sua elevata sensibilità al fotoperiodo. C) Violin plot mostranti la distribuzione dei valori di PCA1 per accessioni Americane (AM) ed Europee (EU) classificate sulla base del gruppo genetico di appartenenza. L'analisi delle componenti principali è stata effettuata sui dati di fioritura (giorni per la fioritura [DTF] e sensibilità al fotoperiodo [PS]) provenienti da 10 esperimenti che includono prove in pieno campo ed in serra in diversi ambienti europei. Il panel C mostra come accessioni americane dei gruppi A2 ed M1 siano sensibili al fotoperiodo (valori elevati di PCA1) ed alcune accessioni Americane della razza Chile siano ancora parzialmente sensibili. C) I grafici a torta mostrano l'assegnazione relativa delle accessioni di ogni gruppo a quel pool genico. Gli individui M2 europei sono in grado di fiorire rispetto alla controparte americana (AM_M2), in quanto gli autori ipotizzano che l'elevato livello di ibridazione e ricombinazione con altre razze abbia permesso l'introgressione di loci per la fioritura insensibili al fotoperiodo. Sottostante i grafici a torta è riportato il numero di accessioni (con relativa percentuale sul totale) caratterizzate da mancata o ritardata fioritura in nord e sud Europa. Immagini tratte da Bellucci et al . (2023), con modifiche.

1.3.2.2 Cece

Figura 11: Immagini rappresentative dello sviluppo e delle caratteristiche fenotipiche nella specie Cicer arietinum. A) piante di cece in fioritura; B) semi di cece; C) fiore di cece.

Il Cece (*Cicer arietinum*) (Figura 11) è una specie erbacea diploide (2n=2x=16; dimensioni del genoma: 740Mpb) (Varshney et al. 2013; Rocchetti et al. 2022) annuale e autogama. I dati FAO riportano che India, Australia e Pakistan rispettivamente guidano la coltivazione mondiale di cece con più di 1 milione di ettari per ciascun paese con una resa media di 1500

Kg/ha che in condizioni irrigue può raggiungere anche 5000Kg/ha (INCREASE). Tra gli stati Europei, sono Spagna, Italia, Bulgaria e Grecia i maggiori produttori di cece.

Vavilov (1926) identificò nel Sudovest Asiatico e Mediterraneo i due centri di origine primari ed uno secondario in Etiopia; mentre Harlan (1992) individuò la mezzaluna fertile come centro di domesticazione. Il processo di domesticazione ha generato un importante effetto "collo di bottiglia" con una riduzione della diversità genetica del pool domesticato rispetto al suo progenitore selvatico (Rocchetti et al. 2022). Il cece coltivato si è diffuso in un primo tempo verso l'Asia e verso il mediterraneo (sud Europa e nord Africa). Nell'ultimo secolo ha infine raggiunto il nord America e l'Australia (Rocchetti et al. 2022).

1.3.2.3 Lenticchia

Figura 12: Immagini rappresentative dello sviluppo e delle caratteristiche fenotipiche nella specie Lens culinaris. A) piante di lenticchia; B) baccelli della lenticchia; C) semi di lenticchia; D) fiori di lenticchia.

La lenticchia (*Lens culinaris*) (Figura 12) è una specie erbacea diploide (2n = 14; dimensione aploide del genoma: 4.3 Gb) annuale ed autogama. La produzione di lenticchia a livello mondiale era di 6.3 Mt nel 2018 con una superficie coltivata di 6.1Mha distribuita in più di 50 stati, tra questi il principale produttore è il Canada (2.1Mt) seguito dall'India (1.6Mt). In Europa sono 241.929 gli ettari coltivati a lenticchia con una produzione totale di 275.565 t e resa media di 11.390 hg/ha (INCREASE).

È stato dimostrato che *L. culinaris* subsp. *Orientalis* è il progenitore selvatico del specie coltivata odierna *L. c.* subsp. *Culinaris* (Liber et al. 2021). La presenza di *L. c.* subsp. *orientalis* è stata dimostrata nel sud-ovest asiatico e raramente in Asia centrale e Cipro (Zohary et al. 2012).

1.3.2.4 Lupino

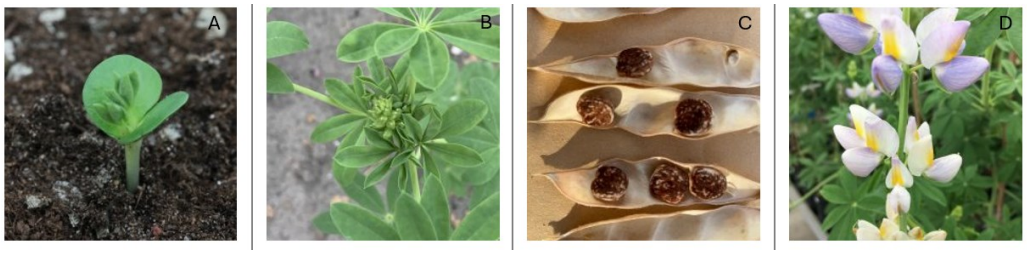


Figura 13: Immagini rappresentative dello sviluppo e delle caratteristiche fenotipiche nella specie Lupinus albus e Lupinus mutabilis. A) plantula da poco germinata di lupino dove possiamo notare i cotiledoni; B) pianta di lupino; C) baccelli di lupino aperti dove possiamo notare i semi; D) fiori del lupino.

Lupinus è un ampio e diversificato genere appartenente alla famiglia delle leguminose, dove possiamo contare fino 1.000 specie appartenenti (INCREASE). Il lupino gode di un'ampia distribuzione geografica (e.g. regioni sub-artiche, regioni mediterranee, regioni montuose dell'africa dell'est, mesoamerica e regioni sub tropicali) il che sottolinea la grande abilità adattativa ai diversi contesti climatici (INCREASE).

Ricerche recenti hanno dimostrato che il genere *Lupinus* ha avuto origine nel "Vecchio Mondo" e successivamente diffuso nel "Nuovo Mondo" (Drummond et al. 2012). Infatti, il genere *Lupinus* è geograficamente distinto in due centri di diversità (Gladstones. 1998), ovvero "Vecchio Mondo" (i.e. bacino del mediterraneo e nord-est Africa) (e.g. *Lupinus albus*) e "Nuovo Mondo" (i.e. nord e sud America) (e.g. *Lupinus mutabilis*) (Figura 13). La domesticazione di *Lupinus albus* (lupino bianco; annuale, 2n=50; dimensione del genoma circa 580 Mbp) risale al tempo dei Greci e dei Romani circa 1000-800 anni A.C., principalmente nell'area della penisola Balcanica e Greca (Gladstones. 1998). Per quanto riguarda *Lupinus mutabilis* (tarwi; lupino Andino; lupino perla; annuale; 2n=48; dimensione del genoma circa 930Mbp) si considera la sua domesticazione nelle Ande tra il 1800 e 2600 A.C. (Atchison et al. 2016)

1.3.3 *Obiettivi del progetto INCREASE e metodologia operativa*

Gli obiettivi del progetto INCREASE prevedono l'identificazione di strategie ottimali per la conservazione, caratterizzazione e conseguente impiego e valorizzazione delle risorse genetiche delle leguminose alimentari. Il progetto coinvolge un'ampia gamma di Stakeholders (i.e. portatori di interesse) come dipartimenti di ricerca e sviluppo, pubblici e privati,

ricercatori, organizzazioni governative e no, scuole e cittadini che interagiranno sotto il coordinamento del progetto INCREASE (Figura 14) (Bellucci et al. 2021), permettendo la produzione di un elevata quantità di dati fenotipici e genotipici (Bellucci et al. 2021). La creazione di nuova conoscenza, ed il coinvolgimento di diversi attori inclusi i consumatori finali e gli agricoltori, promuoverà e migliorerà l'utilizzo sostenibile delle risorse genetiche facilitando l'identificazione dei requisiti e delle necessità degli agricoltori e dei cittadini mediante processi partecipativi (Bellucci et al. 2021).

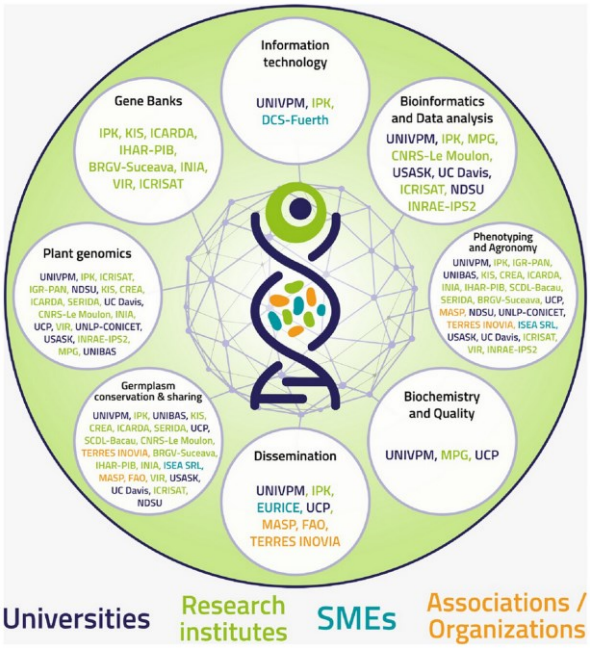


Figura 14:Competenze interdisciplinari e ruoli dei partner del progetto INCREASE. Immagine tratta da Bellucci et al. 2021 con modifiche.

Al fine di esplorare efficientemente la diversità insita nelle risorse genetiche delle colture trattate, INCREASE assemblerà e curerà delle "Collezioni intelligenti" (ICs) come un insieme di Core Collections nidificate di diverse dimensioni in modo da rappresentare l'intera diversità di ogni coltura trattata (Bellucci et al. 2021). Le Collezioni intelligenti sono basate su inbred lines (i.e. prodotte con il metodo Single Seed Descent, SSD) ottenute da un ampio campione di accessioni. Le collezioni intelligenti hanno memoria (i.e. sono basate su inbred lines derivate da accessioni i cui dati fenotipici e genotipici sono noti) e sono capaci di apprendere e integrare, grazie all'analisi dei dati integrati al dataset, informazioni sulla struttura della diversità genetica in relazione al fenotipo e all'ambiente (Bellucci et al. 2021). Questa impostazione fa sì che le Collezioni intelligenti sono capaci di migliorare ed evolvere

correggendo gli errori (anche di campionamento) in base alle nuove informazioni raccolte (Bellucci et al. 2021).

Le collezioni intelligenti saranno così le seguenti (Figura 15):

1) The Reference Core (R-CORE), la collezione più ampia in termini di accessioni e variabile in dimensione per specie. Queste collezioni includeranno 2000-4000 linee SSD che saranno genotipizzate utilizzando un approccio a bassa copertura (e.g., Genotyping by sequencing [GBS]);

2) Il Training Core (T-CORE), che rappresenta un sotto campione dell'R-CORE e comprende circa 400-500 linee SSD per specie. Un approccio di sequenziamento più approfondito è previsto (e.g., Whole Genome Sequencing [WGS]) insieme ad un'ampia caratterizzazione fenotipica classica (supportata da analisi d'immagine) e molecolare sia in condizioni controllate che in campo;

3) L'Hyper-Core (H-CORE), che consisterà in 40-50 linee SSD accuratamente scelte, con l'obiettivo primario di campionare la più grande diversità possibile entro specie, sulle quali verranno effettuate analisi genomiche più approfondite (e.g., WGS, sviluppo di pangenomi) oltre all'approccio di fenotipizzazione già applicato alla T-CORE.

Figura 15: Schema di costituzione delle Increase Intelligent Collections (Bellucci et al., 2021).

Più in dettaglio, l'approccio di fenotipizzazione prevede prove in campo in più località, dove saranno valutati: habitus di crescita, altezza della pianta, sviluppo radicale, efficienza del rapporto simbiotico colore del fiore insieme ad importanti tratti legati alla produzione, resistenza alle fitopatie, giorni necessari alla fioritura, numero e peso di semi e baccelli, etc..,.

Al fine di ottenere tali dati sono applicati protocolli di fenotipizzazione specifici per ciascuna coltura trattata insieme ad analisi di Metabolomica e Trascrittomica (Bellucci et al. 2021). Nell'ambito del progetto sono applicate tecniche innovative al fine di esplorare al meglio il germoplasma disponibile come: "Genome Wide Association Study" (GWAS), "Innovative Machine Learning (ML) Approaches" basato su Intelligenza Artificiale (AI). Le informazioni ottenute a livello fenotipico insieme a quelle genotipiche con le relative predizioni permetteranno il collegamento tra le accessioni caratterizzate e non caratterizzate, comprendere il livello e struttura della diversità delle risorse genetiche e identificare geni di interesse utili per migliorare le capacità adattative e le performance agronomiche delle specie trattate (Bellucci et al. 2021).

1.3.4 *INCREASE "Citizen Science Experiment"*

Nell'ambito del progetto INCREASE è stato implementato un approccio partecipativo di "Citizen Science Experiment" (CSE), nel quale vi è il coinvolgimento di cittadi a livello europeo. I cittadini possono registrarsi mediante una App dedicata (INCREASE CSA) per smartphone, sviluppata nell'ambito del progetto (Bellucci et al. 2021). Una volta registrati, i partecipanti riceveranno direttamente a casa un set di differenti genotipi di fagiolo (cinque Landraces e una varietà di controllo) a partire da 1,082 accessioni della R-Core di INCREASE (Bellucci et al. 2021) (Figura 15).

I cittadini ricevono per posta il materiale relativo all'esperimento, all'interno di una busta postale contenente le accessioni di fagiolo comune debitamente separate in sei bustine contrassegnate dal codice alfa numerico che ne identifica l'accessione. All'interno della busta postale vi sono presenti anche due color Checker (cartoncino millimetrato con ai bordi campioni di colori), documento di benvenuto con istruzioni operative ed un documento recante l'eSMTA (Passaporto).

La circolazione dei semi in Europa è permessa grazie alla generazione di un eSMTA (Standard Material Transfer Agreement) che rappresenta un accordo mutualistico standardizzato per accedere ed utilizzare consapevolmente le risorse genetiche delle piante come indicato da ITPGRFA (Treaty on Plant Genetic Resources for Food and Agricolture) (Bellucci et al. 2021).

Un aspetto interessante dell'esperimento consiste nel fatto che i semi ricevuti dai cittadini vengono poi validati attraverso un sistema di riconoscimento immagine (Bellucci et al. 2021) mediante l'App dedicata, sfruttando il potenziale dell'intelligenza artificiale. Successivamente, i cittadini coinvolti sono coinvolti nella coltivazione dei genotipi ricevuti, in differenti condizioni ambientali, inclusa la semina in campo e in vaso; tale scelta permette di assecondare le disponibilità di spazio dei cittadini coinvolti, rendendo così a portata di mano di ogni cittadino tale esperienza (Bellucci et al. 2021). L'App "INCREASE CSA" dedicata, oltre a permettere la registrazione di dati fenotipici e note sui genotipi ricevuti, contiene dei tutorial sulle varie fasi dell'esperimento: validazione dei semi, disposizione dei genotipi in campo, semina, utilizzo del color Checker, emergenza e riconoscimento delle foglie cotiledonari fino alla metodologia di fenotipizzazione di ciascun carattere richiesto. Inoltre, l'app "INCREASE CSA" permette di condividere le proprie esperienze, utilizzo dei semi raccolti, foto ed opinioni sull'esperimento (Bellucci et al. 2021).

Un punto di forza di questo esperimento risiede nella grande quantità di dati fenotipici raccolti in un numero elevato di ambienti europei, altrimenti non accessibili da pochi ricercatori. Tra questi, ad esempio, dati fenologici (e.g., epoca di fioritura), e morfologici (e.g., caratteristiche dei semi) (nella sezione sperimentale del presente elaborato verranno illustrate con maggior dettaglio il disegno sperimentale, e le procedure di caratterizzazione delle accessioni). I dati fenotipici, inoltre, insieme all'utilizzo di informazioni molecolari, nuove tecnologie nel campo della genomica, ed il ricorso ad approcci di Machine Learning (ML) basati sull'intelligenza artificiale (AI) possono fornire un utile strumento per studiare la variabilità fenotipica dei materiali selezionati. Infine, esperimenti basati su approcci partecipativi possono contribuire a generare maggiore consapevolezza sull'importanza delle risorse genetiche e della loro conservazione e caratterizzazione (Bellucci et al. 2021), oltre a promuoverne la conservazione decentralizzata grazie al supporto di una rete di cittadini. Ulteriori dettagli sull'esperimento di scienza dei cittadini del progetto INCREASE verranno forniti nella sezione di materiali e metodi della parte sperimentale del presente elaborato.

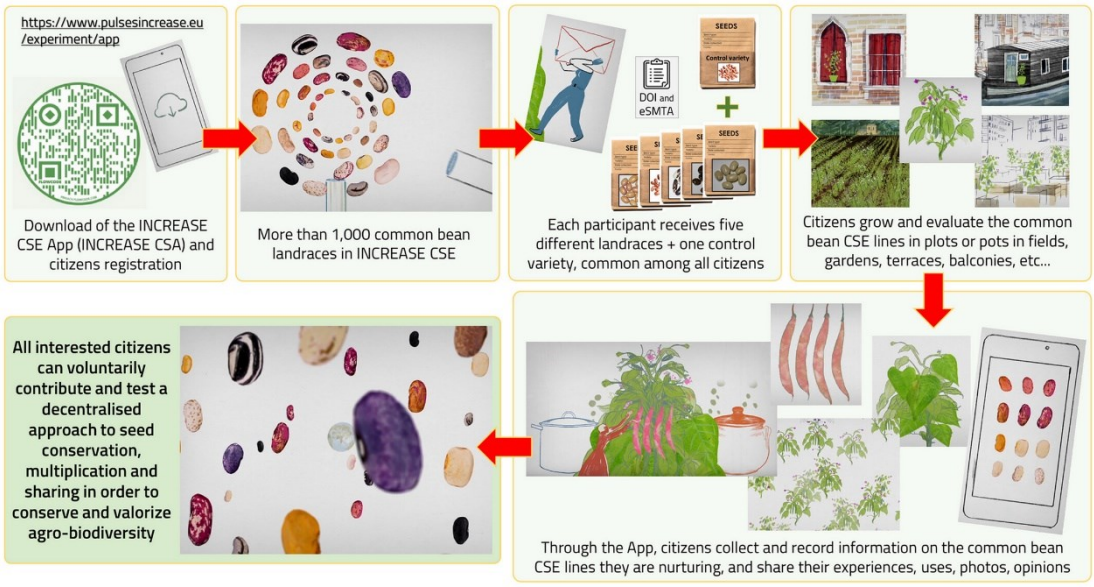


Figura 16:Illustrazione schematica delle singole fasi del Citizen Science Experiment. In ordine orario dall'alto a sinistra; download dell'app "INCREASE CSA"; ricevimento semi fagiolo; validazione, coltivazione, raccolta dati fenotipici e condivisione dei dati e dei semi. Bellucci et al. 2021

1.3.4.1 *Intelligenza artificiale e approcci di Machine Learning per l'analisi di dati provenienti da esperimenti di Scienza partecipata*

Il termine "Intelligenza Artificiale" è stato usato per la prima volta da John McCarthy nel 1956; l'autore la definisce "scienza di creare ed ingegnerizzare macchine intelligenti e in particolar modo programmi informatici intelligenti". M. Sivasubramanian (2021), definisce l'intelligenza artificiale come un campo dell'informatica che ha l'obiettivo di trasmettere l'intelligenza e il pensiero antropico alle macchine e computer in modo che questi possano assistere l'uomo nel proprio lavoro e vita quotidiana. È praticamente impossibile scrivere dei programmi intelligenti, è invece molto più facile e funzionale fare sì che i computer imparino dalle esperienze esattamente come noi umani (Gambus, 2018). Il termine "Machine Learning" infatti consiste nella programmazione di un computer digitale capace di apprendere esattamente come l'uomo (Gambus, 2018).

Questo aspetto viene realizzato mediante l'analisi di un set di dati per il "training" del modello; l'algoritmo, quindi, apprende e diventa capace di riconoscere i pattern delle informazioni in entrata e rendere tali conoscenze generali in modo da rendere possibili previsioni su dati mai visti prima (Alesi. 2022; Nabwire et al. 2021). Tali sistemi aumentano

la qualità delle loro funzioni all'aumentare dei dati raccolti in entrata, in un processo iterativo (Alesi. 2022).

Tutte le tipologie di intelligenza artificiale sono basate sul machine learning che a sua volta trova le fondamenta nelle reti neurali (Figura 17) (Gambus, 2018). Come il machine learning anche le reti neurali sono ispirate dall'uomo; infatti, per la creazione di queste ci si è bastati sul cervello umano dove miliardi di neuroni sono interconnessi tra di loro e processano l'informazione in parallelo (Wang, 2003). In pratica un network neurale artificiale consiste in un livello di input di neuroni (o nodi o unità), uno o due livelli nascosti di neuroni, ed un livello finale di neuroni di output (Wang, 2003).

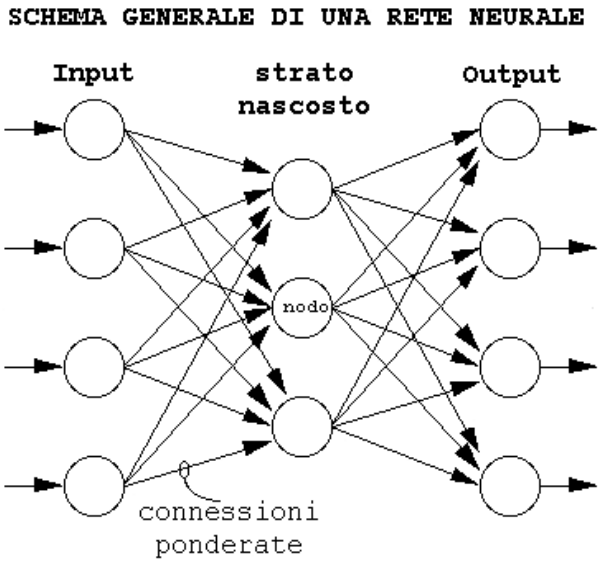


Figura 17: Schema generale di una rete neurale. Immagine tratta da Follador (2008).

È stato dimostrato che così come costruito un network neurale, esso può approssimare ogni funzione computale ad un livello arbitrario di precisione (Wang, 2003). I dati forniti ai neuroni di input sono variabili indipendenti, mentre quelli ottenuti dai neuroni di output sono le variabili dipendenti alla funzione approssimata dal network neurale (Wang, 2003). Questi aspetti offrono interessanti opportunità nel campo del miglioramento genetico; infatti, la fenotipizzazione tradizionale è spesso laboriosa, richiede tempi lunghi ed è a basso rendimento, rendendo così difficile la cattura della natura dinamica delle caratteristiche delle piante (Cembrowska-Lech et al. 2023). Le tecniche di fenotipizzazione ad alto rendimento (Figura 18) vengono attualmente impiegate sia in campo (vedi capitolo "caratterizzazione fenotipica") che in laboratorio o camere di crescita per la raccolta, gestione e analisi dei dati (Nabwire et al. 2021).

Nel campo della fenomica e genomica agraria trova amplia applicazione l'approccio del deep learning (rete neurale con tre o più livelli), che trae vantaggio da ampi dataset utilizzandoli per analizzarne l'immagine mediante l'algoritmo non lineare "Convolutional neural networks" (CNNs) (Lecun et al. 2015). L'applicazione di tali tecnologie nel campo della fenotipizzazione offre la possibilità di ridurre il collo di bottiglia tipico di metodi tradizionali supportando pratiche innovative come l'agricoltura di precisione (Cembrowska-Lech et al. 2023); inoltre, l'integrazione degli approcci di machine learning con le scienze "multi-omiche" (e.g. genomica, trascrittomica, proteomica e metabolomica) potranno fornire una visione olistica del "sistema" pianta, permettendoci quindi di comprendere meglio le complesse interazioni e meccanismi regolatori che caratterizzano le piante (Figura 18) (Cembrowska-Lech et al. 2023).

Un esempio di approccio machine learning per la fenotipizzazione dei tratti morfologici delle piante, lo troviamo nella prova condotta da Falk et al. (2020). In questa prova Falk et al. (2020) hanno sviluppato un protocollo di fenotipizzazione che integra il machine learning con il tradizionale processo di fenotipizzazione delle radici di soia. Il sistema è in grado di integrare l'acquisizione, l'elaborazione e l'analisi delle immagini delle radici ottenute in condizioni controllate fornendo così una piattaforma ad alto rendimento, economica e non distruttiva capace di fornire dati biologici utili per lo sviluppo di studi di genomica e fenomica impiegabili in processi di miglioramento genetico (Falk et al. 2020).

Questi modelli così programmati permetteranno di riconoscere anche altri tratti fenotipici di una coltura (e.g. forma della foglia, colore del fiore etc..) dalla sola immagine e di potenziare strumenti come la Genomic selection (GS) attraverso l'analisi di marcatori dislocati sull'intero genoma (Wenlong Ma et al. 2017) consentendo di sviluppare modelli predittivi circa il fenotipo delle piante (Flores et al. 2023; Biswas et al. 2020; Großkinsky et al. 2018).

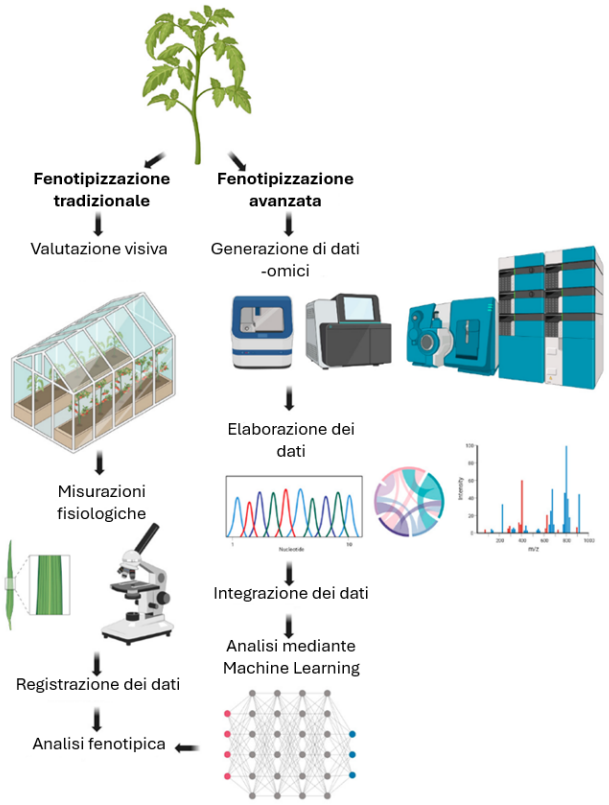


Figura 18: Rappresentazione a confronto delle tecniche di fenotipizzazione; tradizionale (a sinistra) ed avanzata, supportata da AI (a destra). (Cembrowska-Lech et al. 2023 con modifiche).

Il deep learning negli ultimi anni grazie all'evoluzione dei sistemi di calcolo ha rivoluzionato l'analisi delle immagini. In particolare, oggi si distinguono quattro principali aree:

- Classificazione;
- Object detection;
- Image segmentation;
- Instance segmentation.

Nella classificazione, le reti neurali convoluzionali (CNN) vengono addestrate a riconoscere e assegnare etichette a intere immagini, determinando ad esempio se un'immagine rappresenta un cane, un gatto o un'auto. L'attività di object detection, alla base del presente lavoro di tesi, estende questa capacità identificando non solo la presenza ma anche la posizione degli oggetti nelle immagini tramite bounding boxes, permettendo di rilevare più oggetti contemporaneamente. Con la image segmentation si va oltre, suddividendo l'immagine in regioni con significato semantico, assegnando un'etichetta a ogni pixel, utile per applicazioni

come la segmentazione di oggetti differenti all'interno di una scena. Infine, la instance segmentation combina la rilevazione di oggetti con la segmentazione, distinguendo non solo tra diverse classi ma anche tra diverse istanze della stessa classe, identificando singolarmente ogni oggetto come persone in una folla o auto in un parcheggio, fornendo così un livello di dettaglio estremamente elevato.

Tra i vari task sopra delineati l'object detection è la tecnica che ha catturato maggiormente l'attenzione all'interno di tale lavoro di tesi. Una delle tecniche più popolari per l'object detection è YOLO (You Only Look Once) (Redmon et al. 2016). YOLO è basato su reti neurali convoluzionale (CNN) che rileva oggetti e le loro posizioni in una singola passata sulla rete, rendendolo estremamente veloce e adatto per applicazioni in tempo reale.

Il funzionamento di YOLO può essere suddiviso in diverse fasi chiave:

- Suddivisione dell'immagine: L'immagine viene divisa in una griglia di celle. Ogni cella della griglia è responsabile della previsione di un certo numero di bounding boxes e delle loro probabilità di contenere oggetti;
- Predizione dei bounding boxes: Per ogni cella della griglia, YOLO prevede un numero fisso di bounding boxes, ciascuna con le coordinate (x, y, larghezza, altezza) e un punteggio di confidenza che indica la probabilità che la box contenga un oggetto e la precisione della posizione;
- Classificazione degli oggetti: Oltre a predire le bounding boxes, YOLO assegna una probabilità per ogni classe di oggetti per ciascuna bounding box. Questo consente al modello di determinare quale tipo di oggetto è presente in ogni box;
- Soppressione Non-Max (Non-Max Suppression): poiché molti bounding boxes possono sovrapporsi e predire lo stesso oggetto, viene applicata la soppressione non-max per eliminare i box duplicati e mantenere solo quelli con il punteggio di confidenza più alto;

Il vantaggio principale di YOLO è la sua velocità. Poiché l'intera immagine viene processata in una singola passata, YOLO può operare in tempo reale, rendendolo ideale per applicazioni come la videosorveglianza, la guida autonoma e l'analisi dei contenuti video.

Oltre a YOLO, esistono altre tecniche di object detection come SSD (Single Shot MultiBox Detector) e Faster R-CNN (Region-Based Convolutional Neural Networks), ciascuna con i propri vantaggi in termini di velocità e accuratezza. Tuttavia, YOLO rimane uno dei più utilizzati per il suo equilibrio tra precisione e prestazioni in tempo reale (Al Rabbani Alif et al. 2024).

SCOPO DELLA TESI

Nella parte introduttiva del presente elaborato si è discusso dell'importanza della conservazione, caratterizzazione e valorizzazione delle risorse genetiche agrarie come strumento per affrontare le sfide cruciali nel settore agro-alimentare. In particolar modo, si è discusso dell'importanza di valorizzare le risorse genetiche nell'ambito delle specie leguminose, le quali presentano caratteristiche uniche dal punto di vista nutrizionale ed agronomico, al fine di sostenere approcci sostenibili nelle produzioni agrarie ed alimentari. Approcci innovativi, come quelli basati sulla partecipazione di cittadini e stakeholder (i.e., scienza partecipata), possono offrire una soluzione cruciale al fine di studiare l'ampia biodiversità contenuta nelle banche del germoplasma, ad oggi inesplorata, oltre ad offrire un metodo innovativo e dinamico di conservazione delle risorse genetiche agrarie. L'obiettivo del lavoro di tesi è quello di riportare, mediante una serie di dati ed analisi preliminari, la validità di un approccio di conservazione decentralizzata di risorse genetiche nella specie *Phaseolus vulgaris*, basato sul coinvolgimento di cittadini europei. Un esperimento di scienza dei cittadini (CSE; Citizen Science Experiment), nell'ambito del progetto europeo INCREASE (https://www.pulsesincrease.eu/), è stato infatti condotto su oltre 1000 accessioni (linee pure) di fagiolo, distribuite nei primi tre round dell'esperimento (anni 2021-2023) ad oltre 16.798 cittadini Europei, con un picco di partecipanti di 9293 nel terzo Round. Il progetto INCREASE, coordinato dal gruppo di Genetica Agraria del Dipartimento di Scienze Agrarie, Alimentari ed Ambientali dell'Università Politecnica delle Marche si occupa infatti della gestione e caratterizzazione delle risorse genetiche in diverse leguminose alimentari: Cece (*C. arietinum*), Fagiolo comune (*P. vulgaris*), Lenticchia (*L.culinaris*) e Lupino (*L. albus* e *L. mutabilis*), proponendo inoltre strategie alternative e più efficienti per la loro conservazione.

Le risorse genetiche sono fondamentali nei programmi di miglioramento genetico e come fonte di variabilità genetica e fenotipica; tuttavia, spesso il loro potenziale non può essere esplorato, anche per limiti tecnici (e.g., necessità di spazi e risorse per la loro caratterizzazione). Approcci partecipativi come quello identificato nel progetto INCREASE, mediante il CSE, permettono quindi di esplorare la variabilità fenotipica di un numero elevato di accessioni, in un ampio numero di ambienti, consentendo una maggior comprensione della

plasticità fenotipica e degli effetti ambientali su caratteri agronomici rilevanti. Nella sezione sperimentale del presente elaborato verranno illustrate le caratteristiche ed il disegno sperimentale utilizzato nel CSE; inoltre l'obiettivo dell'elaborato è quello di effettuare un'analisi preliminare dei dati registrati dai cittadini, identificando strategie opportune per l'analisi di ampi dataset, con dati provenienti da ambienti differenti, al fine di studiare la qualità del dato registrato e la fattibilità del ricorso ad approcci partecipativi per la caratterizzazione e conservazione decentralizzata delle risorse genetiche agrarie.

Tra i vantaggi del CSE, oltre alla validazione di dati presenti in letteratura e l'ottenimento di nuove informazioni sui materiali studiati, vi è la possibilità di una conservazione decentralizzata delle risorse genetiche (supportata anche dalla possibilità di scambiare materiali tra i cittadini che hanno aderito all'esperimento) e ultimo, ma non meno importante, la possibilità di diffondere un utilizzo consapevole ed incrementare l'interesse verso specie di interesse come le leguminose. Nel presente elaborato, data la disponibilità di ambienti europei molto diversi tra loro per caratteristiche ambientali, verranno analizzati in particolar modo i dati registrati sulla fioritura delle accessioni, in quanto la letteratura riporta ampiamente come cruciali nella transizione verso la fase riproduttiva, fattori ambientali quali temperature e latitudine (sensibilità al fotoperiodo).

Inoltre, i cittadini, oltre a fornire dati, potevano contribuire con l'invio di immagini per caratteri di interesse. Nel presente elaborato sono state catalogate ed etichettate 11.632 immagini ricevute dai cittadini durante lo svolgimento dei primi tre round del CSE, al fine di utilizzarne un sotto set (n=6.960) rappresentanti foglie di fagiolo, per allenare un modello di intelligenza artificiale basato sul machine learning, finalizzato alla predizione di caratteri fenotipici sulla base di immagini. Il carattere "forma della foglia" rappresenterà quindi un modello che verrà poi esteso al riconoscimento di altre caratteristiche come ad esempio, colore del fiore, forma e colore dei semi e caratteri dei baccelli.

Capitolo 1: MATERIALI E METODI

1.1 **Il progetto INCREASE**

Nell'ambito del progetto INCREASE (https://www.pulsesincrease.eu/), coordinato dal gruppo di Genetica Agraria del Dipartimento di Scienze Agrarie, Alimentari ed Ambientali dell'Università Politecnica delle Marche, è stato messo a punto un esperimento partecipativo di caratterizzazione e conservazione delle risorse genetiche mediante il coinvolgimento attivo dei cittadini Europei (i.e., esperimento di Scienza dei Cittadini [CSE]; https://www.pulsesincrease.eu/experiment).

Il progetto mira a definire strategie ottimali per la conservazione, caratterizzazione e conseguente impiego delle risorse genetiche delle leguminose alimentari, fondamentali per le sfide nel campo della sostenibilità agricola, climatica, alimentare e sociale. Le colture trattate nel progetto sono: Cece (*C. arietinum*), Fagiolo comune (*P. vulgaris*), Lenticchia (*L.culinaris*) e Lupino (*L. albus* e *L. mutabilis*). Al fine di esplorare la variabilità insita nelle risorse genetiche delle colture trattate, INCREASE assemblerà delle collezioni intelligenti (ICs) (i.e., basate su Inbred Lines prodotte con il metodo Single Seed Descent, [SSD]) come un insieme di Core Collections nidificate di diverse dimensioni in modo da rappresentare l'intera variabiluitòà della coltura trattata (Bellucci et al. 2021). Attraverso il coinvolgimento di un'ampia gamma di stakeholders e l'attivazione di esperimenti di "Citizen Science" (si veda sezione 1.3.4 del presente elaborato per approfondimenti) il progetto INCREASE è capace di produrre un elevata quantità di dati fenotipici e genotipici (e.g., Whole Genome Sequencing, [WGS]; Genotyping By Sequencing, [GBS]). Al fine di ottenere tali dati vengono applicati protocolli di fenotipizzazione specifici per ciascuna coltura trattata insieme ad analisi di Metabolomica e Trascrittomica. Il progetto prevede inoltre, tecniche innovative per esplorare al meglio il germoplasma disponibile come: Genome Wide Association Study" (GWAS), "Innovative Machine Learning" (ML) ed "Intelligenza Artificiale" (AI). Per ulteriori approfondimenti sul progetto INCREASE si veda introduzione alla sezione 1.3.3.

1.2 L'esperimento di scienza dei cittadini (CSE); Disegno sperimentale e genotipi selezionati

Nell'ambito del CSE, sono state selezionate 1,127 accessioni di fagiolo comune (*P. vulgaris* L.) domesticate provenienti da progetti quali Bean Adapt e Bresov. Le linee sono state scelte in modo da rappresentare efficientemente i due pool genici domesticati della specie (i.e., mesoamericano e andino). Le accessioni selezionate sono prevalentemente landraces (i.e., varietà locali) e cultivar; il 25% delle accessioni presenta habitus nano/cespuglioso mentre il restante 75% presenta habitus rampicante. Le accessioni selezionate sono identificabili mediante un DOI (i.e., Document Object Identifier) ed un codice INCREASE univoco e provengono dalla collezione R-Core del progetto INCREASE (Bellucci et al., 2021). Le accessioni selezionate sono genotipi ottenuti mediante la tecnica del Single Seed Descent (SSD) a partire da un ampio campione di accessioni raccolte dalle banche del germoplasma. Le accessioni SSD ottenute sono state poi moltiplicate in campo e in serra in condizioni controllate in modo da ottenere seme a sufficienza ai fini del CSE. Per queste accessioni sono disponibili dati molecolari (GBS; Genotyping By Sequencing); il 21% del panel selezionato (i.e., 1,082 accessioni provenienti dalla R-core) è in comune con la collezione "nested" T-Core e per queste accessioni si dispone di dati WGS provenienti dal progetto INCREASE (si veda la sezione 1.3.3 del presente elaborato). A queste linee è stata aggiunta una varietà di controllo, la varietà *Meccearly*, selezionata per la fioritura precoce, assenza di sensibilità al fotoperiodo e habitus di crescita determinato e nano.

Le accessioni provengono da oltre 40 diversi paesi (Figura 19).

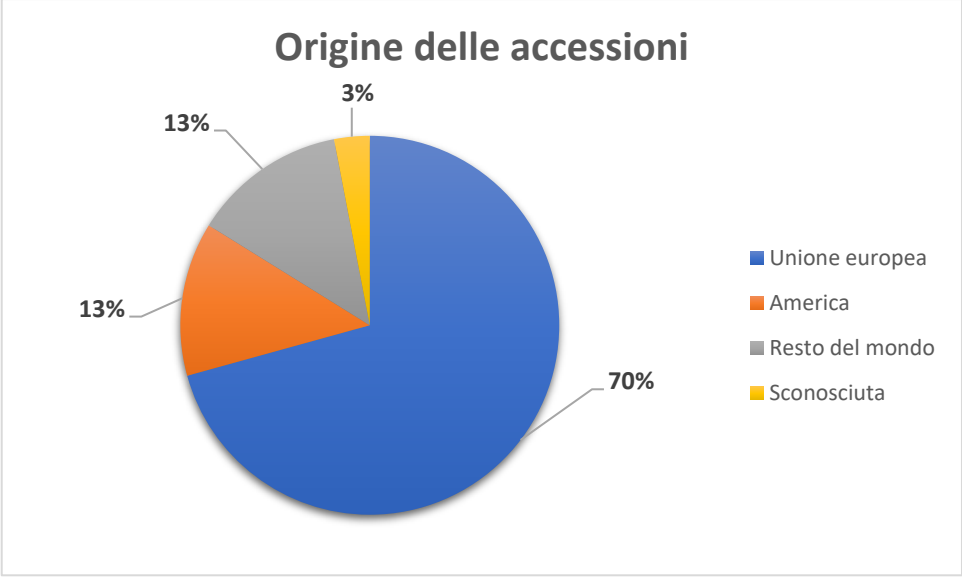


Figura 19: Ripartizione percentuale mediante grafico a torta dell'origine delle accessioni utilizzate nel progetto INCREASE nell'ambito del CSE.

Nell'ambito del CSE, le linee selezionate sono state distribuite a cittadini europei per la caratterizzazione e conservazione decentralizzata, i quali si sono registrati e hanno seguito l'esperimento mediante una App dedicata (si veda la sezione 1.2 del presente elaborato).
Dall'inizio del progetto INCREASE, sono stati effettuati in totale quattro round del CSE (i.e., l'esperimento è stato ripetuto per quattro anni consecutivi a partire dal 2021), di cui uno ancora in corso di svolgimento. A partire dal 2021, che ha visto l'iscrizione di 3,450 cittadini europei, si è potuto notare un sensibile aumento nel numero di cittadini coinvolti fino al 2023 (terzo round del CSE; 9293 iscritti) (Figura 20).

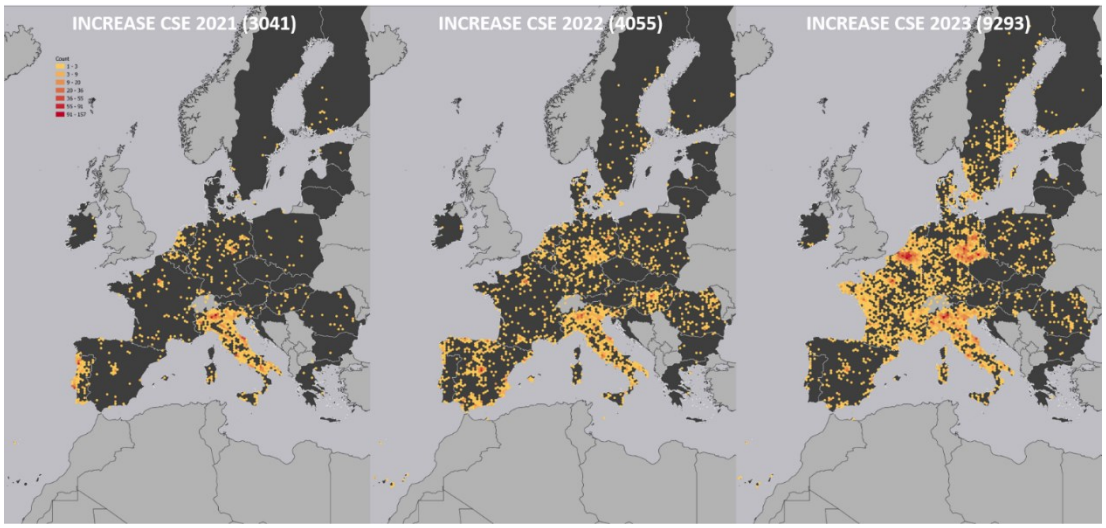


Figura 20:Partecipanti all'esperimento di scienza dei cittadini (CSE) promosso dal progetto INCREASE nel corso dei primi tre round; sinistra) anno 2021, primo round, 3450 partecipanti; centro) anno 2022, secondo round, 4055 partecipanti; destra) anno 2023, terzo round, 9293 partecipanti. Fonte Maps: Markus Oppermann

Nell'ambito dell'esperimento, i cittadini possono collaborare mediante semina delle accessioni in pieno campo o in vaso (ad esempio mediante coltivazione in balcone).
Ogni cittadino iscritto mediante l'App (si veda la sezione 1.2 del presente elaborato) riceve attraverso servizio postale cinque buste contenenti ognuna cinque (in caso di semina in vaso) o dieci (in caso di semina in pieno campo) semi per accessione tra le 1,175 linee, oltre ad una busta contenente il seme della varietà di controllo *Meccearly*. A titolo di esempio, nella preparazione dei semi per lo svolgimento del secondo round del CSE, sono state preparate19,200 buste contenenti semi di accessioni di fagiolo (6,000 buste per semina in vaso e 13,200 per semina in campo).

Come riportato nella sezione 1.3.4 del presente elaborato, insieme alle accessioni di fagiolo comune vengono forniti al cittadino due *color checkers*, passaporto ed un documento riportante le istruzioni operative tradotte nella lingua del paese di residenza del cittadino (Figura 21).

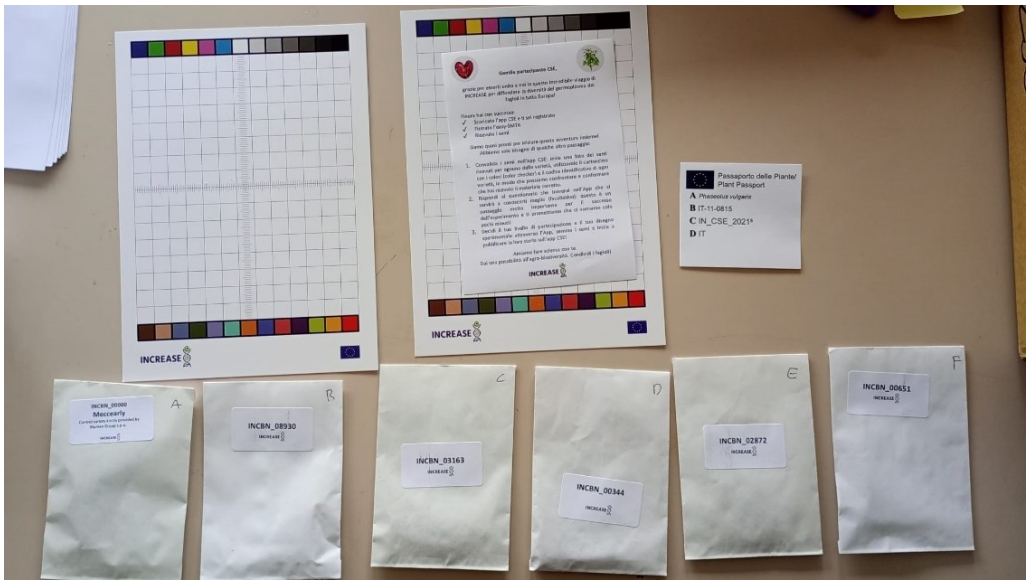


Figura 21: Immagine rappresentante il kit fornito al cittadino per lo svolgimento del CSE.

Ogni linea è stata replicata dalle 10 alle 15 volte. In totale sono stati preparati per la spedizione 3,200 pacchi postali con combinazioni uniche di accessioni di fagiolo comune.
Un disegno sperimentale a blocchi incompleti è stato utilizzato per la randomizzazione delle accessioni. La randomizzazione delle linee del CSE è stata basata su di un set iniziale di 1000 linee, le quali sono state indipendentemente randomizzate in modo da definire dei blocchi di randomizzazione di 200 combinazioni di cinque accessioni per 200 cittadini. In ogni randomizzazione una data accessione è combinata solo una volta con ognuna delle altre accessioni. Il candidato nell'ambito dell'esperimento di CSE ha collaborato attivamente alla preparazione dei semi e del materiale da spedire ai cittadini per i round 3 e 4 (2023 e 2024).

1.3 **INCREASE Citizen Science App (CSA) e raccolta dei dati fenotipici dai cittadini**

L'applicazione "INCREASE CSA" (https://www.pulsesincrease.eu/experiment/app) è fondamentale ai fini del progetto; infatti, attraverso di essa i cittadini possono registrarsi all'esperimento (Figura 22A), accettare l'eSMTA (si veda sezione 1.3.4 del presente elaborato), condividere informazioni, conoscere l'origine delle accessioni coltivate, registrare dati fenotipici (e.g. fioritura, forma della foglia, dimensione del seme etc..) e catturare immagini per documentare i dati relativi ai tratti morfologici (e.g. colore del fiore, forma della foglia, habitus di crescita etc.). Con riferimento alle eSMTA, l'accettazione da parte dei cittadini di uno Standard Material Transfer Agreement è un aspetto cruciale dell'esperimento in quanto la circolazione dei semi in Europa viene garantita attraverso questa procedura.

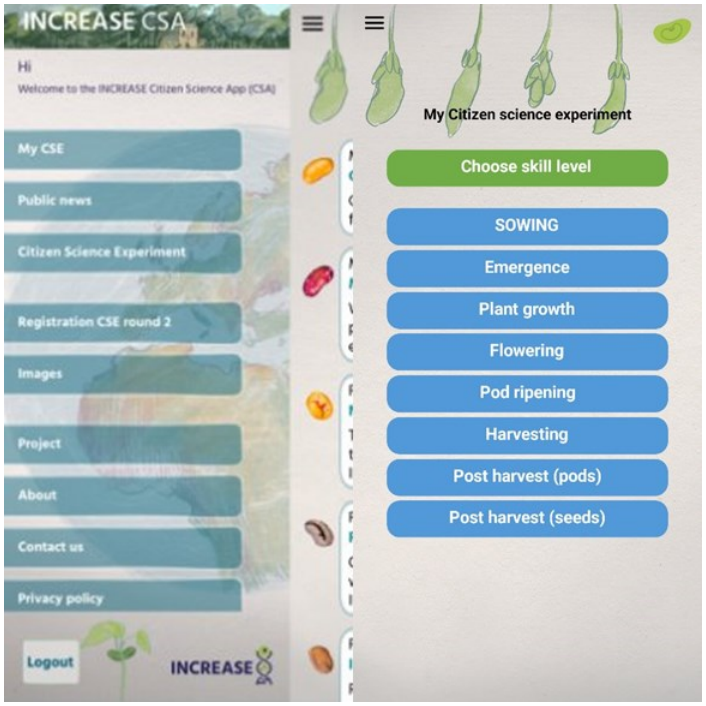


Figura 22: Home Page dell'applicazione "INCREASE CSA" mostrante le varie opzioni di scelta (sinistra), e pagina dedicata al "Citizen science Experiment" (destra).

Il cittadino registrato, e che ha accettato l'eSMTA attraverso l'applicazione riceve il materiale necessario alla conduzione dell'esperimento. Il cittadino procede alla validazione delle accessioni ricevute mediante l'app (Figura 23), questa procedura consiste nello scattare una fotografia dei semi ricevuti, che viene poi confrontata con le fotografie dei semi preparati e spediti dal gruppo di ricerca che coordina il progetto.

Il cittadino può selezionare all'interno dell'App, un protocollo di fenotipizzazione differito sulla base di capacità ed esperienza e distinto a seconda della difficoltà (Figura 22) e Tabella 3).

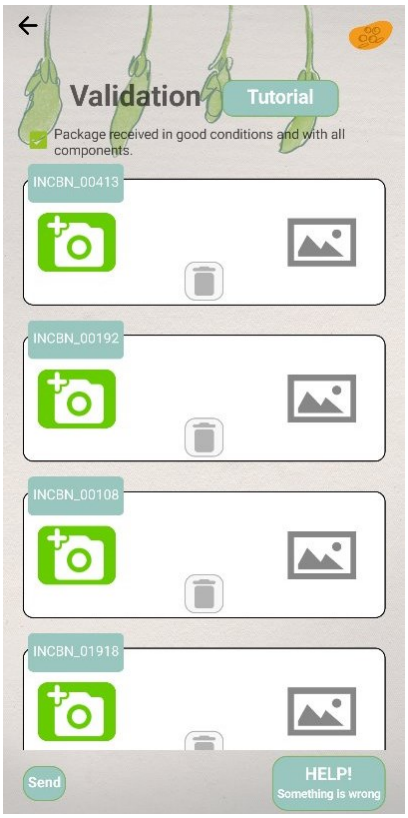


Figura 23: Pagina di validazione delle accessioni ricevute presente nell'applicazione "INCREASE CSA".

I protocolli di fenotipizzazione si distinguono in Base (6 tratti fenotipici da registrare), Medio (21 tratti fenotipici da registrare inclusi i caratteri base) ed Esperto (36 tratti fenotipici da registrare inclusi i caratteri base e medio). La tabella 3 riassume i caratteri sottoposti a fenotipizzazione:

Tabella 3: Caratteri fenotipici soggetti a fenotipizzazione da parte dei cittadini nel progetto di scienza partecipata proposto da INCREASE, distinti in base al livello di esperienza. Il sistema di differenziazione per livelli di difficoltà è stato reso disponibile mediante l'app CSA a partire dal secondo Round. Base, Medio (comprende anche i caratteri base) ed Esperto (comprende anche i caratteri base e medio)

Base	**Medio**	**Esperto**
-Data di semina	-Emergenza	-Pigmentazione dell'ipocotile
-Inizio della fioritura	-Forma della foglia	-Colore delle foglie: clorofilla
-Giorni alla raccolta	-Colore del fiore	-Colore della foglia: antocianina
-Peso totale dei semi	-Inizio della formazione dei baccelli	-Diametro dello stelo
-Numero dei semi	-Curvatura del baccello	-Tipo di crescita
-Morte delle piante	-Colore del baccello	-Massima fioritura
	-Habitus di crescita	-Massima formazione dei baccelli
	-Presenza di fibre nel baccello	-Sezione trasversale del baccello
	-Lunghezza del baccello	-Colore del baccello
	-Larghezza del baccello	-Stato di salute delle piante
	-Tipologia di colorazione del seme	-Numero di baccelli
	-Colore di base del seme	-Lucentezza del seme
	-Colore secondario del tegumento del seme	-Lunghezza del seme
	-Forma del seme	-Altezza del seme
		-Larghezza del seme

Per facilitare e guidare i cittadini nella fenotipizzazione sono stati sviluppati video guida ed esempi nelle varie fasi del procedimento, dalla validazione del materiale fino ai tratti post raccolta (Figura 24 A e B); il protocollo completo di fenotipizzazione è disponibile presso https://www.pulsesincrease.eu/experiment/instructions-and-tutorials.

Per quanto riguarda i caratteri legati alla fioritura nel sito del progetto INCREASE troviamo indicazioni per raccogliere dati in merito ai caratteri "Inizio fioritura", "Massima fioritura" e "Colore del fiore". Per registrare il dato di inizio fioritura occorre, come indicato dalla guida INCREASE, segnare la data di comparsa del primo fiore completamente aperto all'interno della parcella sperimentale (basta da una pianta), attraverso la cattura dell'immagine del fiore con il color checker dietro oppure annotarne la data e caricare il dato sull'App "CSA".

Per registrare il dato di massima fioritura invece, occorre che almeno tutte le piante della parcella sperimentale abbiano un fiore completamente aperto, il dato può essere raccolto attraverso la cattura dell'immagine del fiore con il color checker dietro oppure annotarne la data e caricare il dato sull'App "CSA".

Per quanto riguarda la registrazione del colore del fiore, occorre selezionare all'interno di ogni parcella sperimentale, un fiore rappresentativo proveniente da una pianta ben sviluppata e successivamente registrarne il colore come indicato in figura 24A. la registrazione del dato può avvenire mediante annotazione oppure attraverso la cattura dell'immagine del fiore con dietro il color checker e caricare il dato sull'App "CSA".

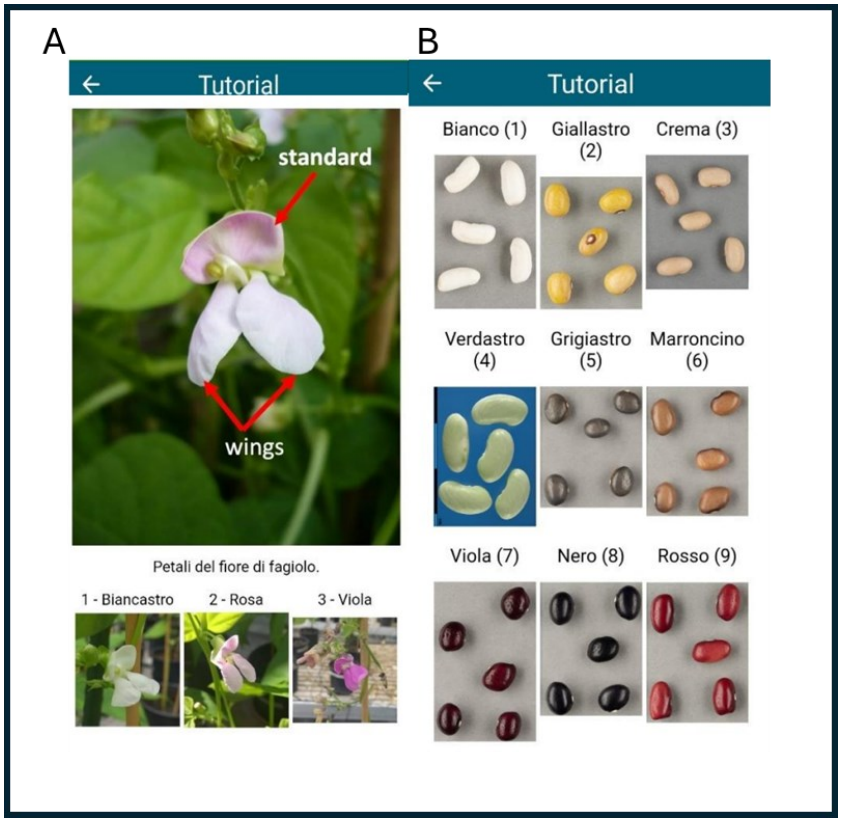


Figura 24: Esempi di tutorial forniti ai cittadini, in merito alla fenotipizzazione dei caratteri richiesti. A) colore del fiore, B) colore del seme.

Ai cittadini viene inoltre richiesto di scattare immagini relative ai tratti fenotipici raccolti, mediante la sovrapposizione dell'oggetto (e.g., foglia, fiore, seme, baccello etc..) sul color checker in condizioni di buona luminosità e stabilità dell'immagine evitando l'interferenza fisica di altri organi della pianta (e.g., sovrapposizioni di foglie, fiori etc.…) (Figura 25).

Le immagini così collezionate dai cittadini possono supportare l'addestramento di un modello di machine learning basato sull'intelligenza artificiale, per analizzare e fenotipizzare i tratti morfologici di interesse (si veda il prossimo paragrafo per maggiori dettagli).

Figura 25: Alcuni esempi di immagini catturate dai cittadini con l'ausilio del color Checker. A) Foglia; B) Fiore; C) Baccello; D) Seme.

I cittadini che hanno partecipato attivamente al CSE potranno inoltre, grazie all'app "INCREASE CSA" scambiare i propri semi ottenuti dalla coltivazione delle accessioni ricevute con altri cittadini partecipanti all' esperimento della scienza del cittadino. Questo passaggio è applicabile ai cittadini che hanno partecipato ai oprecedendi round e che hanno validato con successo i semi delle accessioni ricevute. Lo scambio potrà avvenire mediante una richiesta esplicita tra cittadini (i.e., donatore e ricevente) presentata attraverso l'applicazione "INCREASE CSA" solo dopo l'accettazione dell'eSMTA generato dall'applicazione stessa per il ricevente. Ai cittadini sarà inoltre garantito l'accesso a gruppi Facebook privati moderati dai coordinatori INCREASE dove i cittadini potranno scambiare tra di loro opinioni, esperienze, immagini e consigli

1.4 **Annotazione delle immagini**

La fase di annotazione dei dati è cruciale per l'addestramento efficace di modelli di deep learning per l'object detection, specialmente quando si tratta di riconoscere le forme delle foglie a partire da immagini. Questo processo consiste nel creare un dataset accuratamente etichettato che il modello può utilizzare per apprendere a riconoscere e localizzare diverse forme di foglie. Per ogni immagine, gli annotatori disegnano dei *bounding boxes* intorno alle foglie. Ogni bounding box è definita da quattro coordinate: l'angolo superiore sinistro (x_min, y_min) e l'angolo inferiore destro (x_max, y_max). È importante che i *bounding boxes* siano precisi per rappresentare correttamente la forma delle foglie. Per ogni bounding box deve essere associata a una classe specifica. Per esempio, se si sta annotando un dataset per rilevare foglie lanceolate, ovate e palmato-lobate, ogni box deve essere etichettato con una di queste classi. Dopo l'annotazione iniziale, è importante rivedere il dataset per correggere eventuali errori o inconsistenze.

La qualità della fase di annotazione dei dati è fondamentale per le prestazioni del modello di *object detection* come, ad esempio, YOLO scelto nel caso del presente lavoro di tesi. Annotazioni accurate e consistenti permettono al modello di imparare a rilevare e classificare correttamente le forme delle foglie, migliorando significativamente la sua accuratezza e affidabilità. Di seguito si riportano alcuni applicativi che sono stati utilizzati per effettuare le operazioni di annotazione. Labelbox basato su tecnologie cloud e LabelImg invece che consente di effettuare annotazioni più semplici in locale.

1.4.1 *Labelbox*

Labelbox è un sistema di strumenti di etichettatura che consente di creare dataset pronti per essere utilizzati dagli algoritmi di apprendimento automatico (Figura 26); infatti, molti di questi necessitano di dati di addestramento, i quali devono essere etichettati (Ciaffoni. 2020).

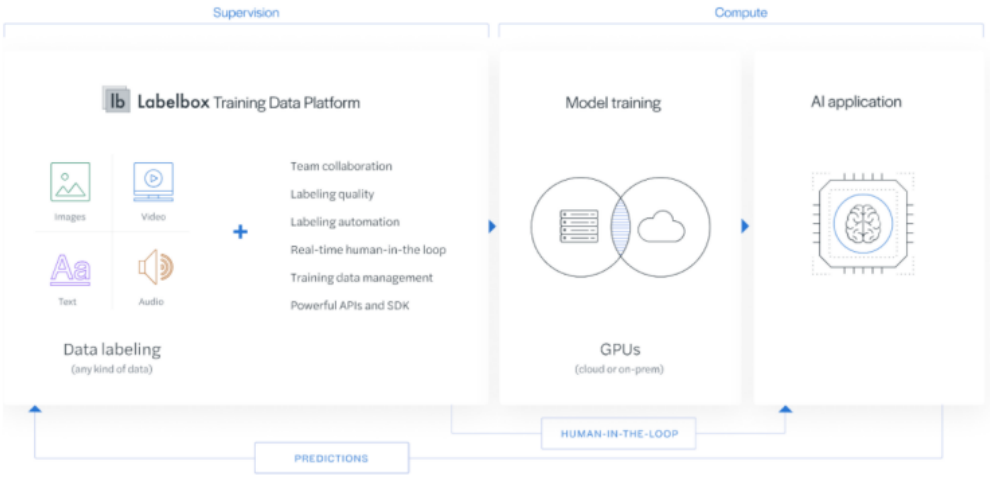


Figura 26: Architettura Labelbox. Ciaffoni 2020

Labelbox presenta diverse opportunità di etichettatura delle immagini; quella adottata in tale situazione è quella del "Bounding Box" (Figura 27), il quale viene utilizzato per etichettare immagini con annotazioni in 2D (Ciaffoni. 2020).

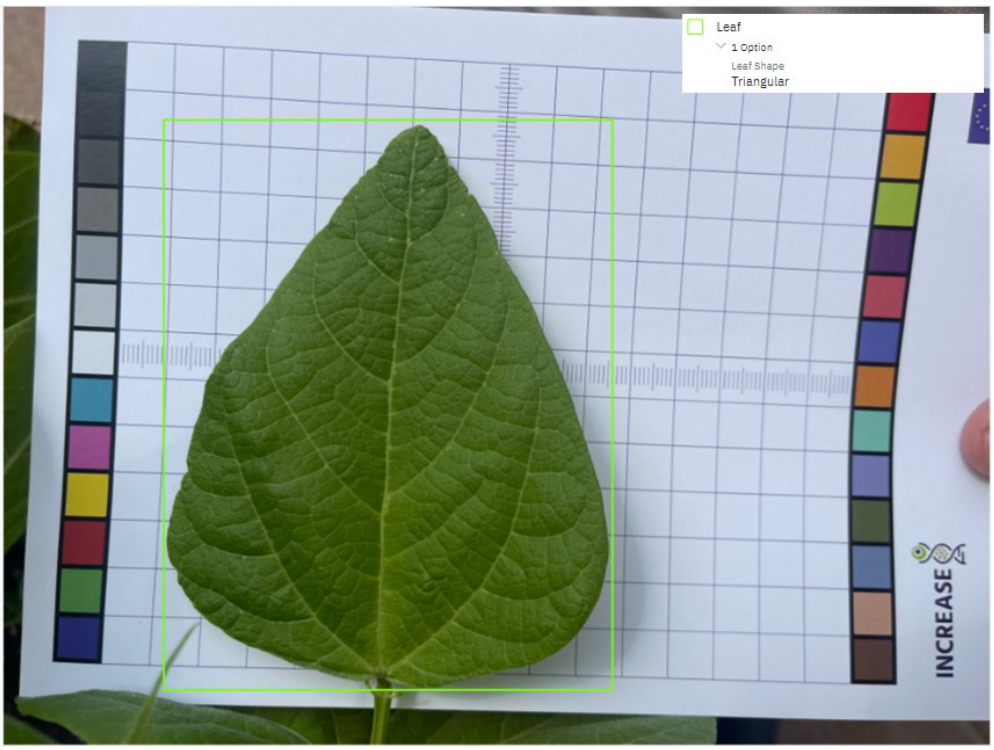


Figura 27: Esempio di etichettatura della forma della foglia mediante il programma "Labelbox". Il rettangolo in verde rappresenta il "Bounding Box"

1.4.2 *Labellmg*

Labellmg (Figura 28) è un programma analogo a Labelbox, utile per l'etichettatura di immagini da dataset per allenare modelli di machine learning. Le annotazioni sono salvate come file XML nel formato PASCAL VOC, ma supporta anche i formati YOLO e CreateML.

Figura 28: Esempio di catalogazione delle immagini mediante il programma "Labellmg". Il rettangolo in verde rappresenta il "Bounding Box", mentre a sinistra è presente la lista delle etichette assegnabili.

1.5 **XML**

Il formato file .XML (eXtensible Markup Language) (Figura 29) è un metalinguaggio sviluppato da World Wide Web Consortium (WC3) per la definizione di linguaggi di markup (Ciaffoni. 2020). L'XML si presenta come linguaggio per creare nuovi linguaggi, atti a descrivere documenti strutturati (Ciaffoni. 2020). Un documento XML si presenta come un file testo che contiene una serie di tag, attributi e testo secondo regole sintattiche ben definite (Ciaffoni. 2020)

```
<annotation>
        <folder>61</folder>
        <filename>96b8700a-75ef-49f3-9cc8-712bc2482f1c.jpg</filename>
        <path>C:\Users\Giaco\Desktop\AI - INCREASE\Sottocartelle immagini CSE_LEAF_SHAPE\61\96b8700a-75ef-49f3-9cc8-712bc2482f1c.jpg</path>
        <source>
                <database>Unknown</database>
        </source>
        <size>
                <width>3120</width>
                <height>4160</height>
                <depth>3</depth>
        </size>
        <segmented>0</segmented>
        <object>
                <name>Triangular</name>
                <pose>Unspecified</pose>
                <truncated>1</truncated>
                <difficult>0</difficult>
                <bndbox>
                        <xmin>1</xmin>
                        <ymin>1208</ymin>
                        <xmax>2381</xmax>
                        <ymax>3845</ymax>
                </bndbox>
        </object>
</annotation>
```

Figura 29: Esempio di file .XML relativo all'etichettatura della forma della foglia.

1.6 **Google Colab**

Google Colab (Figura 30) è una piattaforma che ci permette di eseguire codice direttamente sul Cloud in maniera gratuita, consentendo tra i tanti aspetti di sviluppare sistemi di machine learning dove è necessaria molta potenza di calcolo per addestrare i modelli di rete, soprattutto quando si lavora su dataset di elevate dimensioni (Ciaffoni. 2020). Per eseguire il codice, Google Colab sfrutta il cosiddetto Jupyter Notebook, il quale verrà poi eseguito su macchine virtuali di server Google, consentendo agli sviluppatori di svincolarsi dalla parte hardware e concentrarsi solo sul codice Python e sui contenuti (Ciaffoni. 2020). È possibile configurare la macchina in cui verrà eseguito il nostro codice Python abilitando il supporto all'uso della Graphical Processing Unit oppure supportarsi a configurazioni già esistenti (Ciaffoni. 2020).

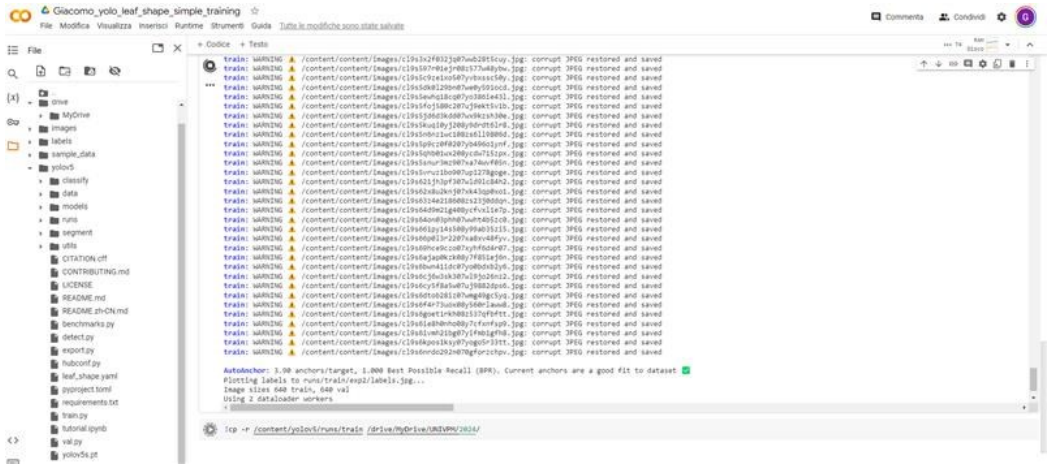


Figura 30: Esempio di schermata di Google Colab relativo al training del modello di machine learning.

1.7 **Etichettatura e costituzione di una libreria di immagini annotate per il carattere forma della foglia**

In un primo tempo ho proceduto con l'annotazione di tutte le 11.632 immagini provenienti dai tre round del CSE attraverso le applicazioni "Labelbox" e "labellmg". Labelbox è stato usato per annotare una serie di immagini con riferimento alla forma delle foglie. Viste le numerose immagini sono stati superati i limiti della piattaforma cloud per account di tipo *free*; per ovviare a questo problema è stato dunque scelto un ulteriore strumento Lablellmg descritto in precedenza. Con tale strumento si è potuto dunque annotare un elevato numero di immagini per il tratto fenotipico in questione (e.g. foglia, fiore, seme e baccello). A ciascuna delle immagini è stato fornito un codice identificatore alfanumerico corrispondente alla relativa accessione INCREASE.

Una volta catalogate le singole immagini è stato scelto come carattere modello la forma della foglia vista l'alta reperibilità delle immagini e semplicità del tratto fenotipico. Questi aspetti sono fondamentali per programmare e testare l'efficacia e affidabilità del modello di machine learning che si intende creare.

Le immagini appartenenti al catalogo delle foglie (circa 11.000) sono state processate grazie ai programmi "Labelbox" e "labellmg" dove ho proceduto in un primo tempo scartando le immagini non idonee e successivamente per le immagini restanti ho etichettato il tratto fenotipico (i.e., foglia) attraverso il tracciamento di un "bounding box" (Figura 31), ovvero un rettangolo che circoscrive l'oggetto etichettato (i.e., foglia). Definito così il "bounding box" ho fenotipizzato la foglia attraverso la selezione di un fenotipo tra i seguenti: triangolare, quadrangolare e rotondo, assegnandolo univocamente all'etichetta come indicato nella figura successiva (Figura 31). Alla fine del procedimento di etichettature le immagini etichettate relative alla forma della foglia sono in totale 6.960.

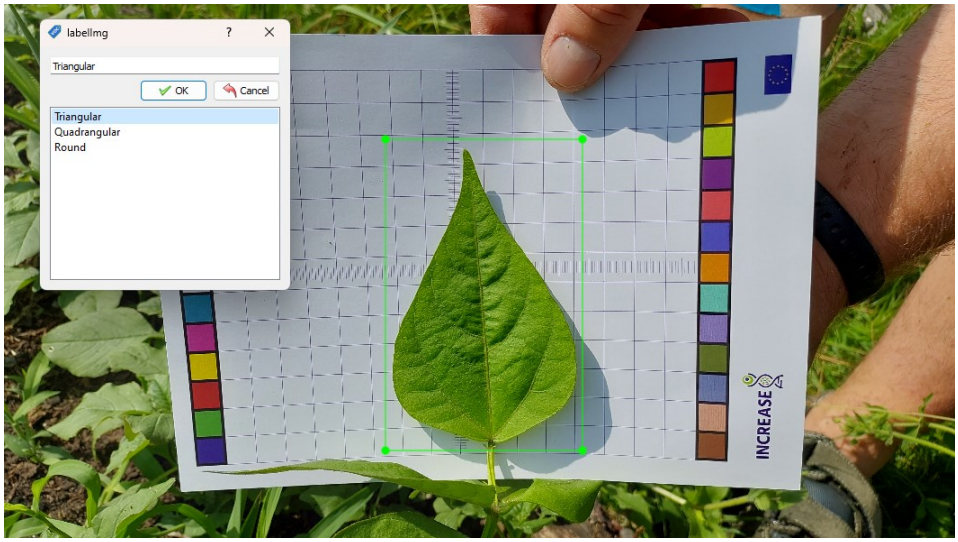


Figura 31: Assegnazione di una etichetta in merito alla forma della foglia mediante "Labellmg".

Il protocollo di fenotipizzazione seguito per l'attribuzione dell'etichetta relativa alla foglia di fagiolo (i.e., triangolare, quadrangolare e rotonda) è stato quello definito da Cortinovis et al. (2021) come indicato dalla seguente figura 32. Questo protocollo viene inoltre seguito, come da tutorial, anche da parte dei cittadini per fenotipizzare la forma della foglia (Tabella 3, tra i caratteri previsti nel livello intermedio).

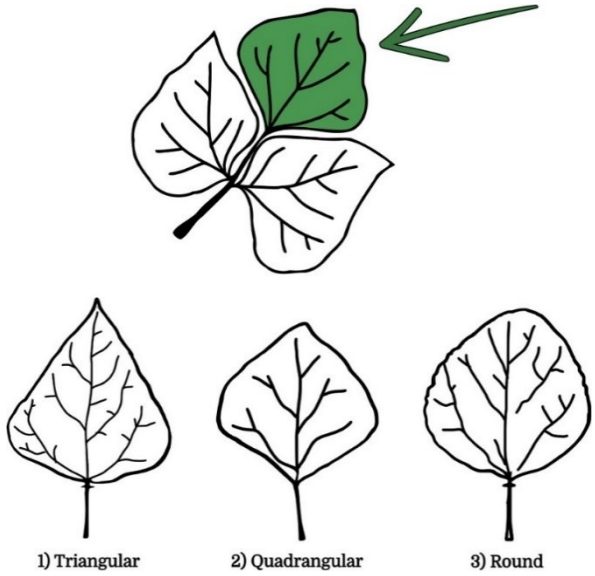


Figura 32: Protocollo di fenotipizzazione proposto per la caratterizzazione della forma della foglia in P. vulgaris. Cortinovis et al. 2021

Grazie all'attribuzione di tutte le etichette alle immagini analizzate, il programma è stato capace di generare un dataset di file .XML riportanti il percorso dei file immagine esaminato con relative informazioni delle etichette generate ovvero disposizione spaziale (x,y) ed il fenotipo attribuito.

Attraverso l'elaborazione di tutti i file .XML generati è stato possibile associare ad ogni accessione il relativo tratto fenotipico per la forma della foglia, come mostrato nella tabella 4.

Tabella 4: Esempio di associazione tra accessione del progetto INCREASE (a sinistra) e fenotipo assegnato alla foglia (a destra).

Accessione	**Tratto fenotipico**
INCBN_00189_CSE_1187.jpg	Quadrangolare
INCBN_00195_CSE_0305.jpg	Triangolare
INCBN_00195_CSE_0306.jpg	Triangolare
INCBN_00198_CSE_0668.jpg	Quadrangolare
INCBN_00198_CSE_2876.jpg	Triangolare
INCBN_00201_CSE_0645.jpg	Rotonda

1.8 **Training, Validazione e Testing**

A valle dell'attività di annotazione (Figura 33), si può procedere all'allenamento di un modello per effettuare il task di object detection. Le fasi di training, validation e testing sono fondamentali nel processo di sviluppo di un modello di object detection come YOLO (You Only Look Once). Si identificano tre fasi principali, ovvero: i) training, ii) validation e iii) testing.

Nella fase di training è cruciale la preparazione del Dataset, ovvero l'insieme delle immagini unitamente alle etichette associate durante la fase di annotazione. In questa fase, il dataset etichettato viene suddiviso in un set di training e un set di validation. Il set di training viene utilizzato per addestrare il modello, mentre il set di validation serve per valutare le prestazioni del modello durante l'addestramento. Successivamente viene configurato il modello, impostando i parametri di addestramento come il learning rate, il batch size e il numero di epoche. Risulta fondamentale anche effettuare delle operazioni preliminari come la fase di *data augmentation*. Per migliorare la robustezza del modello, vengono applicate tecniche di data augmentation come rotazioni, traslazioni, cambiamenti di scala e variazioni di luminosità alle immagini del set di training.

Questo aiuta il modello a generalizzare meglio su immagini non “viste”. Ciò si è reso necessario in quanto alcune classi risultano essere fortemente sbilanciate (si veda immagine seguente).

Figura 33: Esempio di distribuzione delle annotazioni rispetto alle classi di “forma” della foglia (insieme di dati annotati in Labelbox).

La fase immediatamente successiva è quella dell’addestramento del modello. Il modello YOLO viene addestrato utilizzando il set di training attraverso tecniche di transfer learning che consentono di ridurre la complessità dell’apprendimento. Durante l'addestramento, il modello apprende a minimizzare una funzione di perdita che misura l'errore tra le predizioni del modello (bounding boxes e classi) e le annotazioni reali. Questo processo viene iterato per un numero prefissato di epoche. Durante l'addestramento, le prestazioni del modello vengono monitorate utilizzando il set di validation. Metriche come la precisione, il recall e la mAP (mean Average Precision) vengono calcolate per valutare l'accuratezza del modello. In base ai risultati, i parametri di addestramento possono essere ottimizzati.

Di fondamentale importanza è anche la fase di *validation*. Il set di validation viene utilizzato per valutare le prestazioni del modello addestrato. Questo set serve per ottenere una stima di come il modello si comporta su dati non visti durante l'addestramento. In base ai risultati della *validation*, i parametri del modello possono essere aggiustati. Questo processo è iterativo e può includere la regolazione del learning rate, l'aggiunta di ulteriori epoche di addestramento, o modifiche alla struttura della rete neurale. La fase di *validation* aiuta anche a identificare e prevenire l'*overfitting*, assicurandosi che il modello non si adatti troppo strettamente ai dati di training ma generalizzi bene su nuovi dati.

Ultima fase, ma non per questo meno importante, è la fase di testing. Viene utilizzato un set di testing separato e non visto per valutare le prestazioni finali del modello. Questo set deve essere rappresentativo delle condizioni reali in cui il modello verrà utilizzato (es. immagini con ombre o acquisite in situazioni non perfetto come accade delle volte nell’applicazione di CSA Increase. Il modello YOLO viene applicato al set di testing, e le sue predizioni vengono confrontate con le annotazioni reali per calcolare le metriche di prestazione.

Le metriche comunemente usate includono la precisione, il recall e la mAP. I risultati del testing vengono dunque analizzati per valutare le aree di forza e debolezza del modello. Questo può includere l'identificazione di classi di oggetti che il modello riconosce bene e classi su cui il modello ha difficoltà.

È importante continuare a monitorare le prestazioni del modello e aggiornare periodicamente il modello con nuovi dati per mantenere la sua efficacia nel tempo. Queste fasi garantiscono che il modello YOLO sia accuratamente addestrato, validato e testato, assicurando che sia pronto per l'uso in applicazioni di object detection reali con elevate prestazioni.

Capitolo 2: RISULTATI E DISCUSSIONE

2.1 **Partecipazione al CSE e raccolta dei dati**

Nel primo round del CSE, 3,402 cittadini si sono iscritti mediante l'app dedicata CSA; nel secondo round il numero di iscritti è stato di 4,052, con un aumento del 16% rispetto al primo round. Nel terzo round sono state raggiunte 9,340 iscrizioni, con un aumento del 131% rispetto al secondo round. Nel primo round 2,591 dei cittadini iscritti hanno accettato l'eSMTA (76% degli iscritti); 2,988 nel secondo round (74% degli iscritti; con un aumento del 13% rispetto al round 1); 7,302 nel terzo round (78% degli iscritti; con un aumento del 144% rispetto al round 2). I cittadini che hanno successivamente validato le accessioni ricevute sono stati rispettivamente 1,636 nel primo round (48% degli iscritti); 2,988 nel secondo round (55% degli iscritti; con un aumento del 27% rispetto al primo round); 6,220 nel terzo round (67% degli iscritti; con un aumento del 177% rispetto al secondo round). I cittadini che dopo aver validato le accessioni hanno registrato mediante l'app la disposizione sperimentale delle piante in campo sono stati rispettivamente 721 nel primo round (21% degli iscritti); 1,649 nel secondo round (41% degli iscritti; con un aumento del 56% rispetto al primo round); 4,785 nel terzo round (51% degli iscritti; con un aumento del 190% rispetto al secondo round). I cittadini che hanno registrato la data di semina delle accessioni sono stati rispettivamente 704 nel primo round (20% degli iscritti); 1,407 nel secondo round (35% degli iscritti; con un aumento del 50% rispetto al primo round); 4,046 nel terzo round (43% degli iscritti; con un aumento del 188% rispetto al secondo round). I cittadini che hanno registrato la data di emergenza delle accessioni sono stati rispettivamente 563 nel primo round (16% degli iscritti); 555 nel secondo round (14% degli iscritti; con una diminuzione del 1% rispetto al primo round); 1,466 nel terzo round (16% degli iscritti; con un aumento del 164% rispetto al secondo round). La riduzione del dato di emergenza dal primo al secondo round si può spiegare in quanto a partire dal secondo round, con l'introduzione di diversi livelli di difficoltà di fenotipizzazione per i cittadini (Tabella 3) il carattere "emergenza" non risulta più obbligatorio nel livello Base. I cittadini che hanno registrato la data di fioritura delle accessioni sono stati rispettivamente 262 nel primo round (8% degli iscritti); 620 nel secondo round (15% degli iscritti; con un aumento del 58% rispetto al primo round); 2,108 nel terzo round (23% degli iscritti; con un aumento

del 240% rispetto al secondo round). I cittadini che hanno registrato la data di formazione del baccello sono stati rispettivamente 188 nel primo round (5% degli iscritti); 198 nel secondo round (5% degli iscritti; con un aumento del 5% rispetto al primo round); 696 nel terzo round (7% degli iscritti; con un aumento del 252% rispetto al secondo round). Infine, I cittadini che hanno registrato la data di raccolta dei semi delle accessioni sono stati rispettivamente 93 nel primo round (3% degli iscritti); 302 nel secondo round (7% degli iscritti; con un aumento del 69% rispetto al primo round); 1,675 nel terzo round (18% degli iscritti; con un aumento del 455% rispetto al secondo round).

Tabella 5: Statistiche relative al CSE (Round 1,2 e 3) indicanti il numero di cittadini coinvolti nelle varie fasi dell'esperimento. Si evidenziano le percentuali dei cittadini partecipanti rispetto al totale, nelle singole fasi, e la variazione percentuale di tali valori in confronto al round precedente.

	Statistiche round 1		**Statistiche round 2**			**Statistiche round 3**		
	Numero cittadini	**% sul totale dei cittadini registrati**	**Numero cittadini**	**% sul totale dei cittadini registrati**	**Variazione % rispetto round 1**	**Numero cittadini**	**% sul totale dei cittadini registrati**	**Variazione % rispetto round 2**
Registrazione all'esperimento mediante l'App	3402	100	4052	100	+16	9340	100	+131
Accettazione eSMTA	2591	76	2988	74	+13	7302	78	+144
Validazione accessioni	1636	48	2247	55	+27	6220	67	+177
Registrazione disegno sperimentale	721	21	1649	41	+56	4785	51	+190
Registrazione data semina	704	21	1407	35	+50	4046	43	+188
Registrazione data emergenza	563	17	555	14	-1	1466	16	+164
Registrazione data fioritura	262	8	620	15	+58	2108	23	+240
Registrazione data allegagione	188	6	198	5	+5	696	7	+252
Registrazione raccolta	93	3	302	7	+69	1675	18	+455

Con riferimento all'analisi dei dati, è necessario considerare che la numerosità dei cittadini coinvolti offre, da un lato, la possibilità di investigare la variabilità fenotipica entro un ampio set di accessioni studiate, in una moltitudine di ambienti potenzialmente differenti tra loro. Al fine di analizzare i dati tenendo conto della variabilità ambientale, diverse strategie sono state testate.

Per l'analisi dei dati, con riferimento al secondo round del CSE sono state considerate 12 unità di randomizzazione (RUs) a livello europeo (Figura 34), in grado di coprire diversi areali geografici, nei quali erano localizzati i cittadini iscritti, sia con prove di campo che in vaso. Le randomizzazioni a livello geografico, per gli esperimenti in campo dei cittadini sono state classificate sulla base di variabili come: latitudine, altitudine e tre variabili bioclimatiche (Bio 10: media temperatura trimestre più caldo; Bio 18: precipitazioni trimestre più caldo; Bio 3 isotermalità). La classificazione è riassunta nella tabella 6.

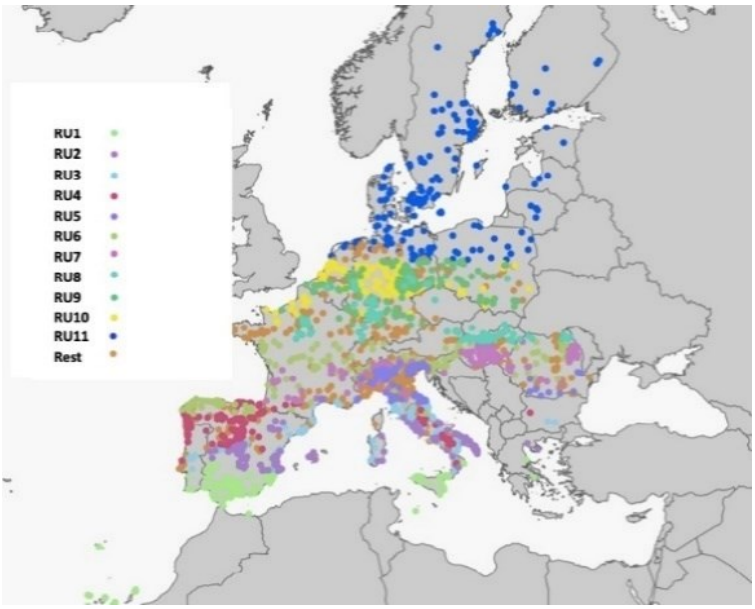


Figura 34: Distribuzione delle RUs nel territorio Europeo adottata nel secondo round di CSE; a sinistra la legenda con la RUs ed il relativo colore identificatore.

Tabella 6: Distribuzione delle RUs e corrispondenza con paesi europei (a destra) in base alle caratteristiche ambientali (latitudine, temperatura media e altitudine). Sistema adottato durante il secondo round del CSE.

RU	Caratteristiche ambientali	Paesi
1	Latitudine (36.6); Temperatura media (24.4 °C)	Spagna, Italia, Portogallo, Grecia
2	Latitudine (40.2); Temperatura media (23.9 °C)	Spagna, Italia, Grecia
3	Latitudine (41.3); Temperatura media (22.6 °C)	Italia, Spagna, Portogallo, Francia, Bulgaria
4	Latitudine (42.1); Temperatura media (19.4 °C); Elevata altitudine	Spagna, Italia, Portogallo, Francia, Bulgaria
5	Latitudine (45.1); Temperatura media (22.1 °C)	Italia, Francia, Romania
6	Latitudine (45.4); Temperatura media (18.5 °C); Elevata altitudine	Spagna, Francia, Italia, Romania, Slovenia, Ungheria, Svizzera, Austria
7	Latitudine (46.0); Temperatura media (20.1 °C)	Ungheria, Italia, Romania, Francia, Slovacchia
8	Latitudine (49.0); Temperatura media (19.0 °C); Bassa altitudine	Germania, Ungheria, Francia, Romania, Austria, Slovacchia, Polonia, Svizzera, Belgio, Repubblica ceca
9	Latitudine (50.8); Temperatura media (17.4 °C)	Germania, Polonia, Francia, Belgio
10	Latitudine (51.2); Temperatura media (16.5 °C)	Germania, Olanda, Francia, Polonia, Belgio
11	Latitudine (56.5); Temperatura media (16.0 °C); Bassa altitudine	Svezia, Germania, Danimarca, Polonia, Finlandia, Olanda, Lettonia, Lituania, Estonia

Osservando le date di registrazione per caratteri e step cruciali dell'esperimento, nell'ambito del Round 2 del CSE, si può osservare una netta scalarità dovuta alle differenti condizioni ambientali in cui i cittadini conducono l'esperimento (Figura 35A e 35B). Ad esempio, in media, la semina è avvenuta in anticipo nelle RU1-5 (che corrispondono a regioni con temperature medie superiori e latitudini inferiori), rispetto alle RU8-11, caratterizzate da temperature medie inferiori. Le date di semina scalari, e le diverse condizioni ambientali si riflettono in diverse epoche in cui i cittadini delle diverse RU registrano dati fenologici, come ad esempio l'epoca di fioritura (Figura 35A e 35B).

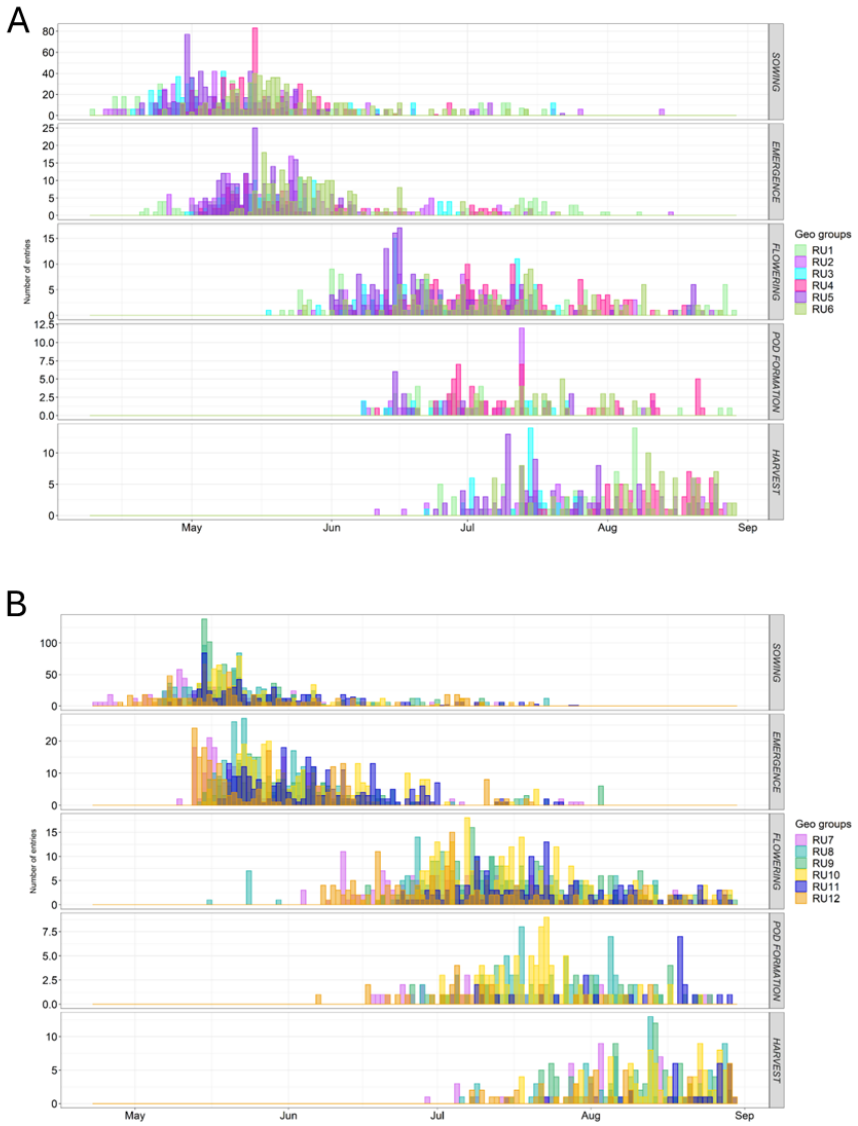


Figura 35: Distribuzione della registrazione dei dati (Semina, Emergenza, Fioritura, Allegagione, Raccolta) da parte dei cittadini (Round 2), per Unità di Randomizzazione geografica (RU). A) dati relativi alle RUs1-6. B) dati relativi alle RUs7-12. Si veda Figura 34 e Tabella 6 per informazioni sulla distribuzione delle randomizzazioni.

Nell'ambito del Round 2 del CSE, confrontando i dati raccolti per la varietà di controllo (*Meccearly*), disponibile in ogni ambiente/ cittadino e le rimanenti accessioni tra differenti RU (Figura 36), si osserva che, mentre non vi sono significative differenze nell'epoca di emergenza, probabilmente a causa di una semina contemporanea del controllo e delle altre accessioni da parte dello stesso cittadino, in tutte le RU, la varietà di controllo si conferma precoce, rispetto alla media delle altre accessioni.

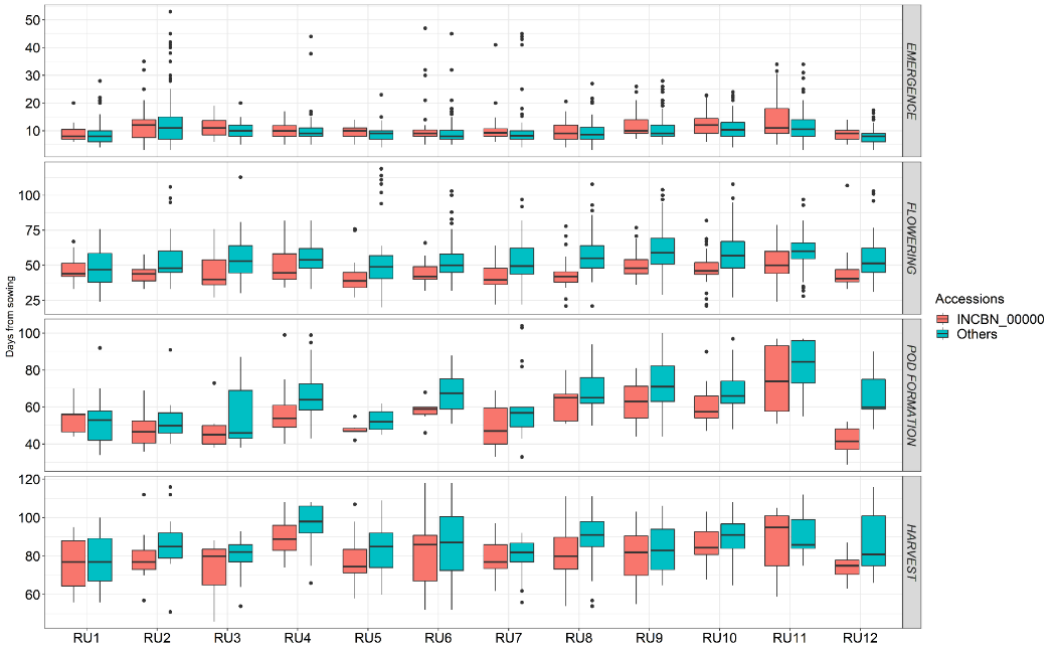


Figura 36: Box plot mostrante la variabilità relativa ai dati raccolti dai cittadini (Semina, Emergenza, Fioritura, Allegagione, Raccolta) per ogni RUs nel corso del secondo round di CSE. I dati fanno riferimento al controllo (in rosso) altre accessioni (Landraces; in blu).

La variazione del numero di cittadini iscritti tra round diversi del CSE rende necessario adattare il disegno sperimentale e la strategia per l'analisi dei dati disponibili. Con riferimento alla suddivisione in aree geografiche e unità di randomizzazione (RUs), l'elevato numero di cittadini iscritti nel terzo round, corrispondenti ad ambienti spesso molto diversi tra loro, rende problematica l'identificazione di un numero di RUs rappresentative di macrogruppi; si è quindi optato per una suddivisione che tenga conto principalmente di fattori ambientali quali la latitudine e la temperatura, considerati cruciali per la transizione da fase vegetativa a riproduttiva. Per questo nelle analisi esplorative fatte finora abbiamo rappresentato il dato solo come nord, centro e sud EU. L'analisi dei dati nel Round 3 è stata condotta adottando una suddivisione dei cittadini in tre unità di randomizzazione sulla base della latitudine (nord Europa, centro Europa e sud Europa) (Figura 37). I dati raccolti dai cittadini in merito alle fasi di fenotipizzazione (i.e., semina, emergenza, fioritura, formazione del baccello e raccolta) sono caratterizzati da una diversa distribuzione temporale della registrazione del dato a seconda della macroarea geografica individuata sulla base della latitudine (Figura 37). In linea generale, i cittadini classificati come parte del Sud Europa hanno avviato e registrato precocemente le fasi dell'esperimento rispetto ai cittadini del Centro e Nord Europa.

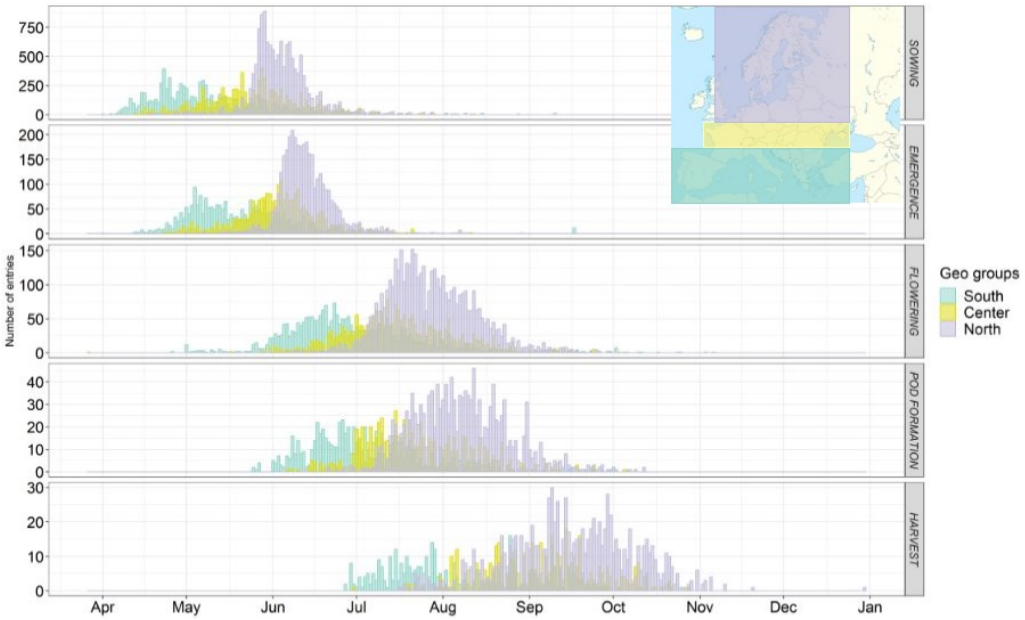


Figura 37: Distribuzione della registrazione dei dati (Semina, Emergenza, Fioritura, Allegagione, Raccolta) da parte dei cittadini per il terzo round di CSE. I dati sono stati registrati dai cittadini e vengono raggruppati per gruppi geografici; Nord Europa in viola, Centro Europa in giallo, Sud Europa in celeste.

Con riferimento ai dati registrati durante il terzo Round del CSE, si osserva che per importanti caratteri fenologici e agronomici, la varietà di controllo *Meccearly*, nota per essere particolarmente precoce, raggiunge in media le fasi di fioritura, allegagione e raccolta prima delle altre accessioni, in tutte le aree geografiche considerate (Figura 38).
I dati presentati confermano la possibilità di utilizzare i dati registrati dai cittadini, in quanto si mostrano in linea con l'atteso.

Inoltre, con riferimento all'analisi dei dati di fioritura, di cui si discuterà nel successivo paragrafo, oltre ad esprimere il dato per ogni accessione come "giorni alla fioritura dalla semina", i dati sulla precocità e stabilità del controllo (Figura 38-Fioritura; 39) supportano la possibilità di normalizzare i dati di fioritura di ogni accessione utilizzando il dato di fioritura (giorni dalla semina) del controllo, entro singolo cittadino/ambiente. Quest'ultimo metodo di normalizzazione offre la possibilità di mitigare le differenze in termini di data di semina tra ambienti/cittadini diversi, in quanto si rapporta il dato di ogni accessione su una varietà di controllo comune.

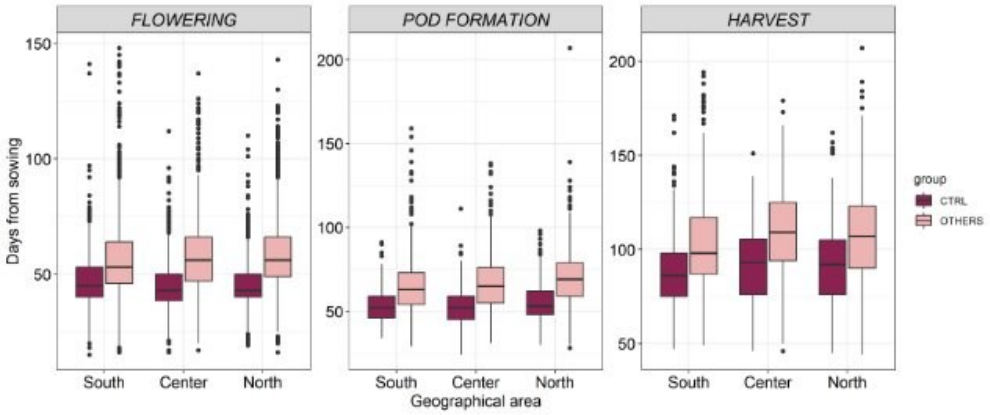


Figura 38: Box plot mostrante la distribuzione in termini di giorni dalla semina fino a diversi stadi fenologici (fioritura, allegagione e raccolta) della varietà controllo e delle altre accessioni nel terzo round di CSE su base geografica (Sud, Centro e Nord Europa).

2.2 **Dati di fioritura registrati dai cittadini**

Il dato "days to flowering (DTF)" (giorni per la fioritura) registrato dai cittadini è espresso come giorni per raggiungere lo stadio di primo fiore dalla data di semina. L'analisi del dato tiene conto quindi di questo sistema di registrazione per esprimere il dato di fioritura. Inoltre, sulla base dei dati di fioritura della linea di controllo (*Meccearly)*, che ne confermano la precocità, si è deciso di normalizzare i dati di fioritura di ogni singola accessione utilizzando il dato di fioritura della linea di controllo, entro ogni ambiente/cittadino. In figura 39, si riportano i dati di fioritura nel secondo round del CSE, della sola linea di controllo, nelle varie RUs, così come descritte nel paragrafo 2.1 (Tabella 6). In particolare, si può osservare come la maggior precocità venga riscontrata soprattutto in regioni a clima più mite (RU5; Figura 39). L'utilizzo di un controllo precoce, stabile e condiviso in tutti gli ambienti, al fine di normalizzare la data di fioritura delle altre accessioni, può consentire una migliore analisi del dato, confrontando dati di fioritura provenienti da ambienti/ cittadini differenti dove le semine sono effettuate in modo scalare e non uniforme. Il dato di fioritura delle altre accessioni viene quindi espresso come numero di giorni che precedono o seguono la fioritura dell'accessione controllo, potendo quindi assumere anche valori negativi, nel caso (seppur raro) in cui una accessione fiorisca prima del controllo. È importante sottolineare, con riferimento all'analisi dei dati, che trattandosi di dati registrati da cittadini, aventi diversi background e competenze, ci si potrebbe aspettare un dato non sempre accurato come quando registrato da uno scienziato. Tuttavia, come si evidenzierà nel presente elaborato, i dati registrati sono in linea con quanto ci si potrebbe attendere, sulla base delle informazioni più recenti disponibili in letteratura.

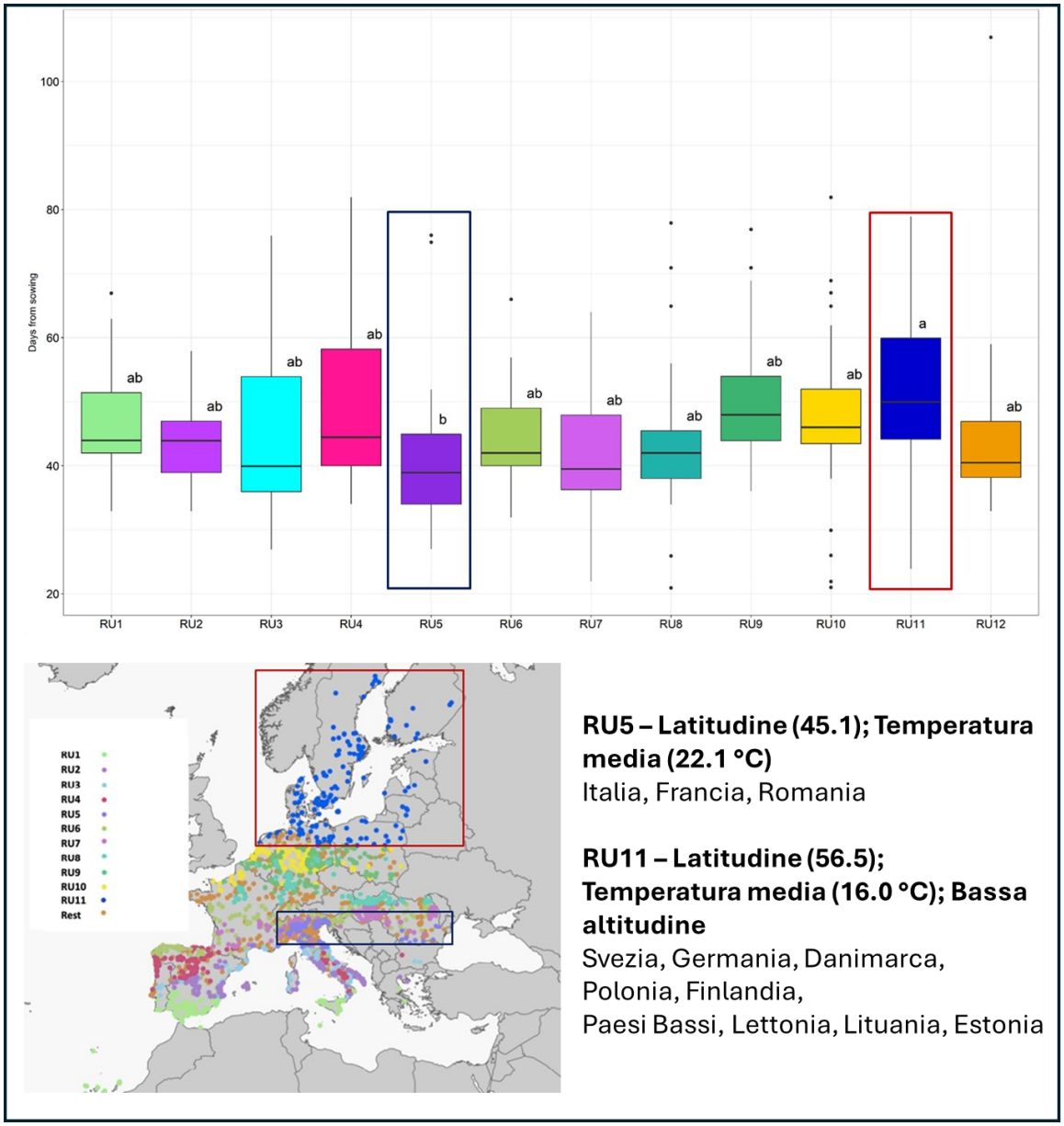


Figura 39: Dati di fioritura (secondo round del CSE) della varietà di controllo nelle RUs. A) Box plot mostrante la distribuzione della variabilità in termini di giorni per la fioritura raggruppata per RUs nella varietà controllo (Meccearly) nel secondo round di CSE. Nella RU5 si osserva la maggior precocità di fioritura come media dei cittadini facenti parte di quella RU. Nella RU11 la linea di controllo fiorisce più tardivamente rispetto alle altre RU (pur confermandosi come linea precoce, in termini generali, se confrontata con altre accessioni studiate). B) Distribuzione delle RUs nel territorio Europeo; a sinistra la legenda con la RUs ed il relativo colore identificatore.

Nell'ambito del progetto INCREASE, come riportato nel paragrafo 1.3.3 del presente elaborato, uno degli obiettivi è quello di produrre dati molecolari (WGS e GBS) per le accessioni impiegate in vari esperimenti.

Per quanto riguarda i materiali studiati nel CSE, nell'ambito del progetto (analisi non facente parte del presente elaborato) sono stati ottenuti dati di GBS (genotyping by sequencing) su 968 accessioni; ciò ha permesso l'identificazione, a seguito di opportuni filtraggi, di 1770 SNPs (Single Nucleotide Polymorphism) per lo studio della diversità genetica del panel di accessioni di fagiolo del CSE. Attualmente è in corso il completamento dell'analisi GBS sull'intero panel del CSE. Un'analisi ADMIXTURE (Alexander et al. 2009) con questi dati, ha permesso di comprendere la struttura genetica delle accessioni del CSE, ed in linea con quanto già descritto nel lavoro di Bellucci et al. (2023) (vedi paragrafo introduttivo 1.3.2.1), nell'ambito di INCREASE è stata confermata la corrispondenza tra specifici gruppi genetici e le razze di fagiolo comune descritte da Singh et a. (1991), con l'identificazione di accessioni appartenenti prevalentemente ai gruppi genetici A1 (razza Nueva Granada), A2 (Razza Peru), A3 (Razza Chile), M1 (Razze Durango e Jalisco), M2 (razza Mesoamerica) (Figura 40).

Come possiamo osservare dalla Figura 40A e 40B, che riporta i dati di fioritura del Round 1 del CSE per le accessioni raggruppate per gruppi genetici, , le accessioni del pool genico A1 (Nueva Granada) sono tendenzialmente le più precoci rispetto alle altre razze, mentre le accessioni appartenenti al pool genico A2 (Perù) sono mediamente più tardive. I dati a disposizione, seppur limitati per alcuni gruppi genetici, confermano quanto osservato in letteratura (Bellucci et al., 2023) sui dati di fioritura nei pool genici di fagiolo. Infatti, le accessioni della razza Perù (A2) mostravano difficoltà di adattamento e mancata diffusione in Europa a causa della sensibilità al fotoperiodo (Bellucci et al., 2023). Inoltre, il confronto delle Figure 40A e 40B, confermano come l'utilizzo di un controllo precoce (Varietà *Meccearly*) per la normalizzazione dei dati contribuisca a ridurre la dispersione del dato intorno alla media.

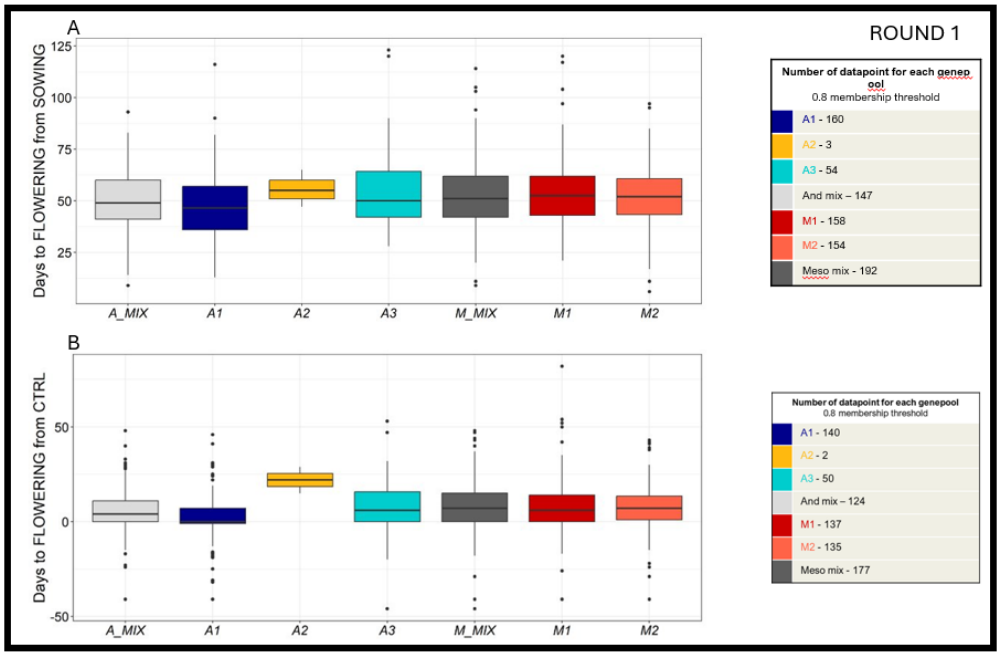


Figura 40: Dati di fioritura (Round 1 del CSE) delle accessioni raggruppate per gruppi genetici. A) Box plot mostrante la distribuzione dei giorni alla fioritura dalla semina nei diversi pool genici individuati da Bellucci et al. (2023) nel primo round di CSE, a destra il numero di individui per ogni pool genico. B) Box plot mostrante la distribuzione dei giorni alla fioritura normalizzato sulla base della data di fioritura del controllo nei diversi pool genici individuati da Bellucci et al. (2023) nel primo round di CSE, a destra il numero di individui per ogni pool genico. M_mix: insieme di accessioni Mesoamericane delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Mesoamericana; A_mix: insieme di accessioni Andine delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Andina.

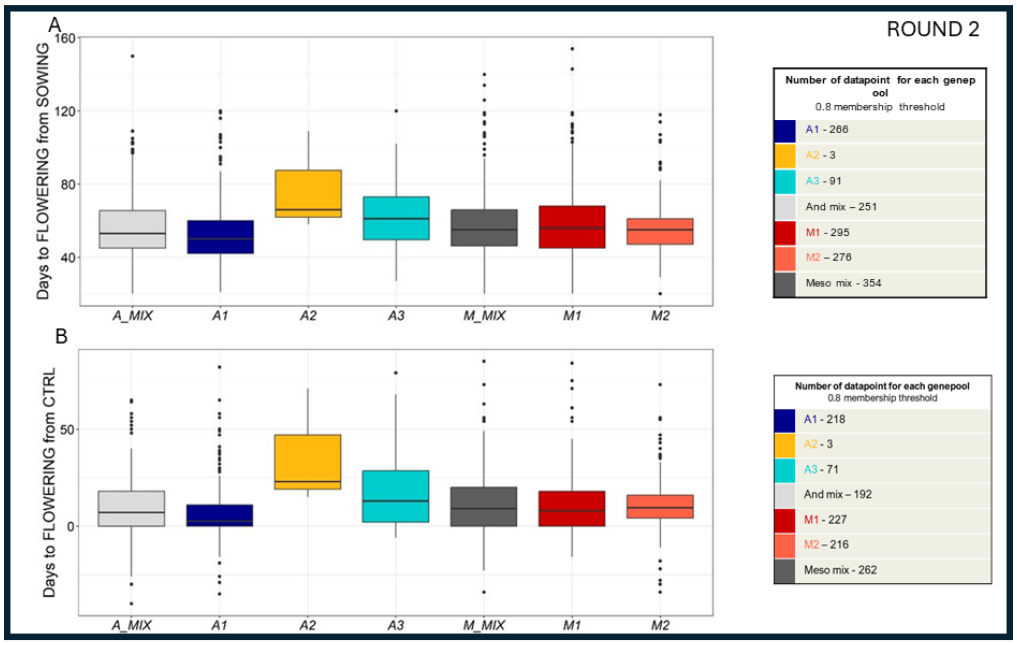


Figura 41: Dati di fioritura (Round 2 del CSE) delle accessioni raggruppate per gruppi genetici. A) Box plot mostrante la distribuzione dei giorni alla fioritura dalla semina nei diversi pool genici individuati da Bellucci et al. (2023) nel secondo round di CSE, a destra il numero di individui per ogni pool genico. B) Box plot mostrante la distribuzione dei giorni alla fioritura normalizzato sulla base della data di fioritura del controllo nei diversi pool genici individuati da Bellucci et al. (2023) nel secondo round di CSE, a destra il numero di individui per ogni pool genico. M_mix: insieme di accessioni Mesoamericane delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Mesoamericana; A_mix: insieme di accessioni Andine delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Andina.

Il secondo Round del CSE permette di disporre di un numero maggiore di dati per alcuni dei gruppi genetici considerati, data la maggior partecipazione dei cittadini. I dati di fioritura relativi al secondo round di CSE (Figura 41A e 41B) conferma quanto osservato nel primo round, con la maggior precocità del gruppo genetico A1 e la tardività delle accessioni del pool genico A2 (Figura 41A e 41B). Con riferimento ai dati di fioritura del secondo round, per le accessioni del pool genico A3 (razza Chile) si può notare inoltre in media una minor precocità nella fioritura rispetto ad altri gruppi genetici, seppur non in modo netto come osservato per il gruppo A2.

Questi dati sono in linea con quanto descritto da Bellucci et al. (2023); gli autori riportano infatti la presenza di una certa sensibilità al fotoperiodo nelle accessioni del gruppo genetico A3, che mostrano problemi di fioritura quando allevate in esperimenti di pieno campo nel nord Europa rispetto al Sud Europa. Sebbene il numero di dati a disposizione non ci permetta una rigorosa analisi statistica, quello che osserviamo è in linea con la letteratura e indica come dei dati presi da cittadini diano risultati promettenti per analisi future

I dati a disposizione nel Round 2 confermano anche l'assenza di sensibilità al fotoperiodo nelle accessioni dei gruppi genetici M1 e M2. .

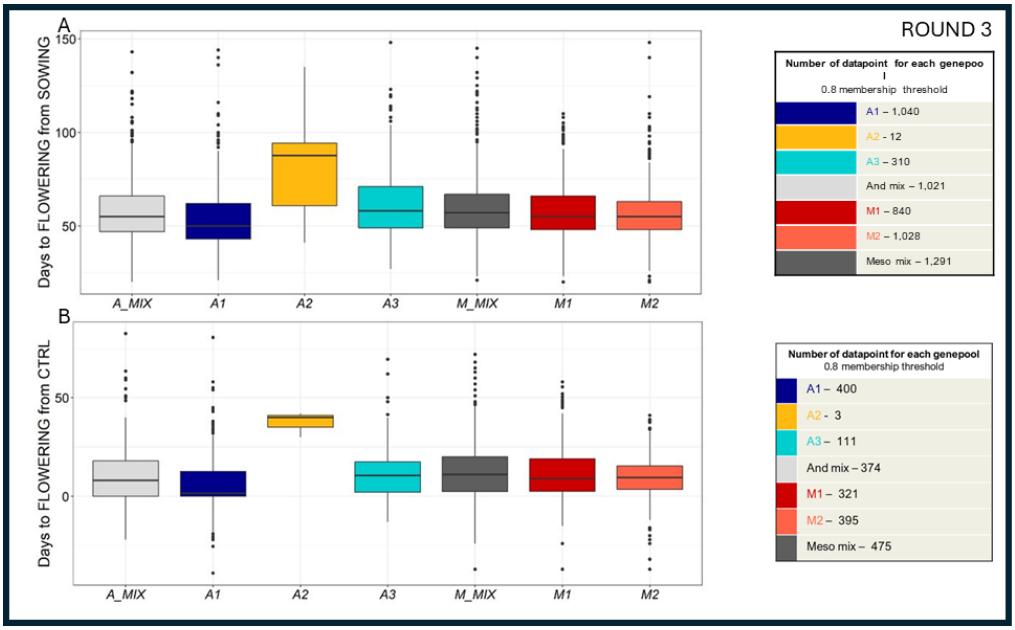


Figura 42: Dati di fioritura (Round 3 del CSE) delle accessioni raggruppate per gruppi genetici. A) Box plot mostrante la distribuzione dei giorni alla fioritura dalla semina nei diversi pool genici individuati da Bellucci et al. (2023) nel terzo round di CSE, a destra il numero di individui per ogni pool genico. B) Box plot mostrante la distribuzione dei giorni alla fioritura normalizzato sulla base della data di fioritura del controllo nei diversi pool genici individuati da Bellucci et al. (2023) nel terzo round di CSE, a destra il numero di individui per ogni pool genico. M_mix: insieme di accessioni Mesoamericane delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Mesoamericana; A_mix: insieme di accessioni Andine delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Andina.

I dati del terzo Round, per il quale si è registrato il maggior numero di iscritti e cittadini che hanno registrato il dato di fioritura (Tabella 5), sono in linea con i risultati dei precedenti Round, mostrando la maggior precocità delle accessioni A1, in particolar modo rispetto al gruppo genetico sensibile al fotoperiodo A2 (Figura 42A). Tali differenze sono evidenti quando si ricorre all'utilizzo della varietà di controllo come strumento per la normalizzazione dei dati di fioritura (Figura 42B). Osservando i dati relativi ai giorni per la fioritura dalla semina (Round 3; Figura 43A) di ogni pool genico nelle tre aree geografiche europee (Sud, Centro e Nord Europa) (si veda Figura 43B), si può osservare come ad eccezione delle accessioni A1 (Nueva Granada), la media dei giorni alla fioritura aumenti spostandosi partire dal sud verso il nord Europa (i.e., all'aumentare della latitudine) (Figura 43B). Per quanto riguarda il pool genico A2 (Perù) non si osservano fioriture nel nord Europa (Figura 43A). Inoltre, a fioritura osservata per le accessioni del gruppo genetico A2 (Perù) si conferma tardiva in particolar modo per i cittadini del centro Europa (Figura 43A), confermando quanto già riportato ed i dati disponibili in letteratura (Bellucci et al., 2023).

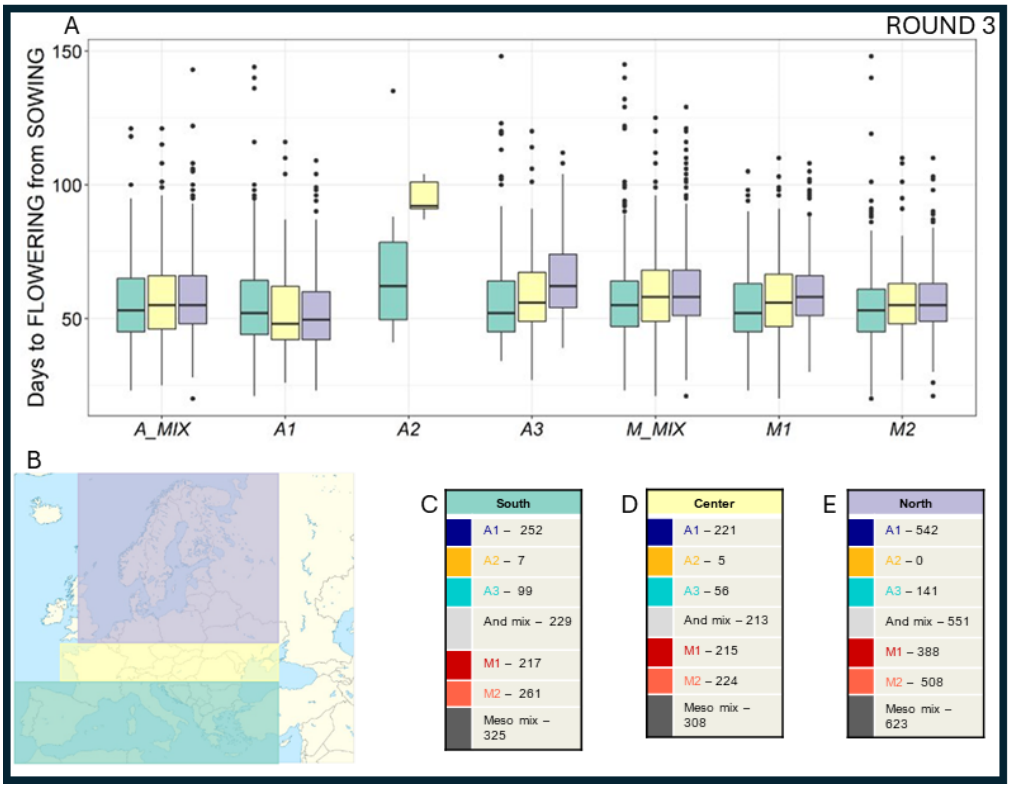


Figura 43: Dati di fioritura dalla data di semina (Round 3 del CSE) delle accessioni raggruppate per gruppi genetici e sulla base dell'origine geografica del dato (Cittadini del Sud, Centro e Nord Europa). A) Box plot mostrante la distribuzione della variabilità in termini di giorni per la fioritura dalla semina dei pool genici individuati da Bellucci et al.

(2023) per il terzo round di CSE. Per ogni pool genico sono state distinte le accessioni provenienti dal nord, centro e sud Europa. M_mix: insieme di accessioni Mesoamericane delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Mesoamericana; A_mix: insieme di accessioni Andine delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Andina. B) Distribuzione dei gruppi geografici distinti in Nord Europa in viola, Centro Europa in giallo, Sud Europa in celeste. C) Numero di accessioni afferenti al gruppo geografico "sud Europa" distinte in base ai pool genici di provenienza. D) Numero di accessioni afferenti al gruppo geografico "centro Europa" distinte in base ai pool genici di provenienza. E) Numero di accessioni afferenti al gruppo geografico "nord Europa" distinte in base ai pool genici di provenienza.

Il ricorso alla normalizzazione dei dati sulla base della data di fioritura della varietà di controllo permette di confermare quanto osservato precedentemente (Figura 44A), ed i risultati di Bellucci et al. (2023) sulla sensibilità al fotoperiodo, capacità di adattamento e diffusione dei gruppi genetici in seguito alla loro introduzione in Europa, che può essere considerata come un centro di domesticazione secondario. I dati inoltre dimostrano come il ricorso ad approcci partecipativi (e.g., esperimenti di Scienza dei Cittadini) sia una strategia di conservazione decentralizzata e di caratterizzazione delle risorse genetiche efficace e affidabile, in quanto permette di replicare dati riportati in letteratura.

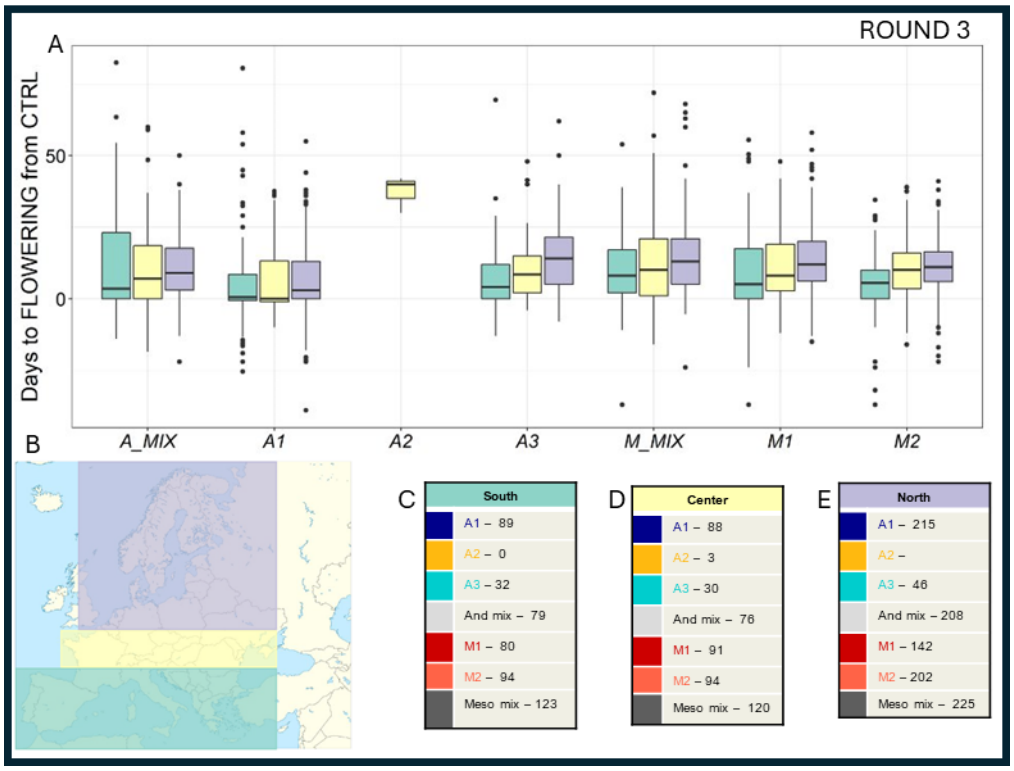


Figura 44: Dati di fioritura normalizzati sulla varietà di controllo (Round 3 del CSE) delle accessioni raggruppate per gruppi genetici e sulla base dell'origine geografica del dato (Cittadini del Sud, Centro e Nord Europa). A) Box plot mostrante la distribuzione della variabilità in termini di giorni per la fioritura normalizzato sulla base della data di fioritura della varietà controllo dei pool genici individuati da Bellucci et al. (2023) per il terzo round di CSE. Per ogni pool genico sono state distinte le accessioni provenienti dal nord, centro e sud Europa. Nel caso specifico di A2 mancano i Box Plot di sud e nord Europa per via della mancata fioritura della varietà controllo utilizzata per normalizzare i dati. M_mix: insieme di accessioni Mesoamericane delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Mesoamericana; A_mix: insieme di accessioni Andine delle quali non si conosce l'esatta appartenenza ad un preciso pool genico di origine Andina. B) Distribuzione dei gruppi geografici distinti in Nord Europa in viola, Centro Europa in giallo, Sud Europa in celeste. C) Numero di accessioni afferenti al gruppo geografico "sud Europa" distinte in base ai pool genici di provenienza. D) Numero di accessioni afferenti al gruppo geografico "centro Europa" distinte in base ai pool genici di provenienza. E) Numero di accessioni afferenti al gruppo geografico "nord Europa" distinte in base ai pool genici di provenienza.

2.3 **Risultati ottenuti dagli algoritmi di intelligenza artificiale per il riconoscimento della forma della foglia.**

Prima di introdurre i risultati ottenuti dagli algoritmi di intelligenza artificiale per il riconoscimento della forma della foglia è necessario introdurre uno strumento di base che è necessario per la comprensione dei risultati, ovvero il concetto di matrice di confusione. Tale strumento fondamentale nell'ambito della classificazione ed è utilizzato per valutare le performance di un modello di classificazione. Nel caso di una matrice di confusione multi-classe, essa si estende al contesto in cui ci sono più di due classi da prevedere. Queste metriche sono cruciali per comprendere la performance di un modello di classificazione multi-classe e per identificare eventuali problemi di bilanciamento tra le classi.

2.3.1 *Matrice di Confusione Multi-Classe*

Una matrice di confusione multi-classe è una tabella di dimensioni n×n, dove n è il numero delle classi. Ogni riga della matrice rappresenta le istanze della classe predetta, mentre ogni colonna rappresenta le istanze della classe reale.

Ad esempio, per una classificazione con tre classi (A, B, C), la matrice di confusione potrebbe apparire così:

Tabella 7: Esempio di matrice di confusione.

	Classe A Predetta	**Classe B Predetta**	**Classe C Predetta**
Classe A Reale	50	2	1
Classe B Reale	5	45	3
Classe C Reale	2	1	52

2.3.1.1 Overall Accuracy

L'accuratezza generale (Overall Accuracy) è una misura globale. Si calcola come il rapporto tra il numero di previsioni corrette e il numero totale delle previsioni.

$$\text{Overall Accuracy} = \frac{\sum_{i=1}^{n} \text{true positives}_i}{\text{Total Samples}}$$

Figura 45: Formula impiegata per il calcolo della "Overall Accuracy".

Dove il numeratore rappresenta la somma degli elementi diagonali della matrice di confusione ed al denominatore il numero totale degli esempi; quanto più il valore è alto (max 1) tanto più il modello “globalmente” ha performato bene.

2.3.1.2 User Accuracy (Accuratezza dell'Utente)

La user accuracy, o precisione, per una classe specifica è il rapporto tra i veri positivi (True Positives) per quella classe e il totale delle previsioni fatte per quella classe (somma degli elementi della riga corrispondente).

2.3.1.3 Producer's Accuracy

La producer accuruacy, o recall, per una classe specifica è il rapporto tra i veri positivi per quella classe e il totale delle istanze reali di quella classe (somma degli elementi della colonna corrispondente).

2.3.2 *Preparazione dei dati*

Un aspetto particolarmente importante è stato quello di applicare tecniche di data augmentation per incremenare il nunero di immagini per alcune classi poco rappresentante e per garantire una maggiore robustezza del modello sviluppato. Le tecniche permettono di aumentare artificialmente il numero di immagini di addestramento, migliorando la generalizzazione del modello. Di seguito si riportano alcune delle tecniche di data augmentation più comuni e utili per l'identificazione delle foglie e delle loro forme: i) rotazione, ii) traslazione, iii) zoom, iv) flipping (vertical/horizontal), v) cambi di luminosità. Nella seguente figura si mostra l’applicazione delle tecniche di augmentation ad una immagine acquisita nel contesto del progetto.

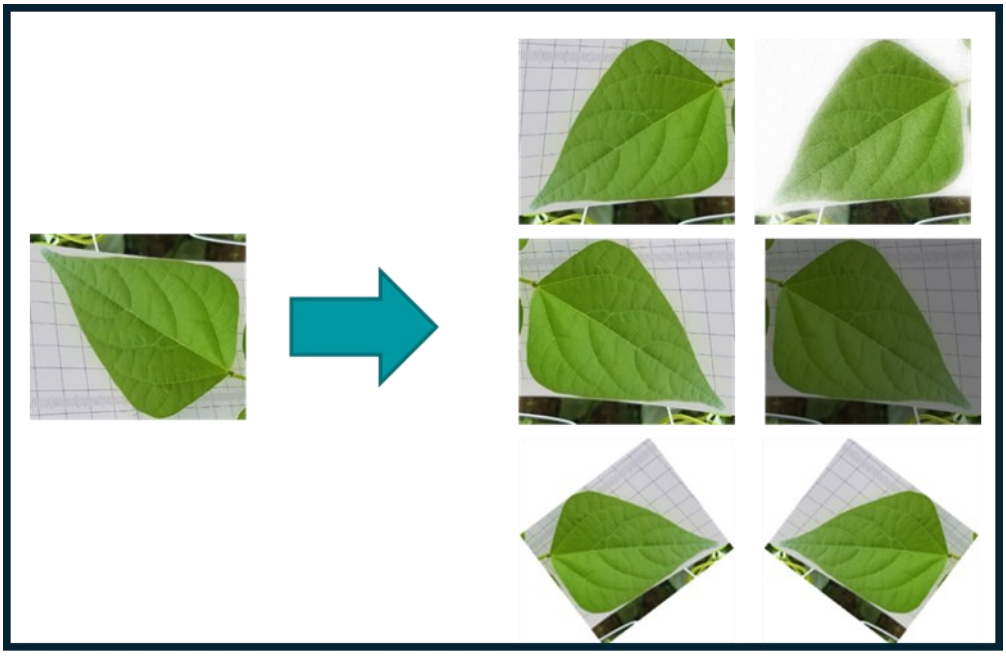


Figura 46: Applicazione delle tecniche di data augmentation ad una immagine.

2.3.3 *Training e testing del modello*

2.3.3.1 Da annotazione a data-set

Dopo aver utilizzato uno strumento di annotazione per etichettare le immagini è necessario esportare le etichette in formato YOLO, con un file *.txt per ogni immagine (se non ci sono oggetti nell'immagine, non è necessario alcun file *.txt). Per tale scopo sono stati creati degli script python al file di passare dai formati di annotazione degli strumenti descritti nei paragrafi precedenti (Labelbox e LabelImg) al formato richiesto ovvero YOLO. I formati di ingresso erano rispettamente JSON ed XML Le specifiche di file *.txt sono:

- Una riga per oggetto
- Ogni riga è in formato classe x_centro y_centro larghezza altezza.
- Le coordinate dei bounding box devono essere in formato x y w h normalizzato (con valori da 0 a 1).

2.3.3.2 Estrazione delle immagini da usare per training – validation e test.

A partire dalle immagini raccolte, dopo un processo di selezione sono state selezionate un sotto-insieme di immagini ed in particolare di seguito si dettagliano i due data-set, dove T, Q, R sono rispettivamente le etichette associate alle classi Triangular, Quadrangular, Round.

- Dataset #1 (11/2023)
    - T 2386
    - Q 390
    - R 55
- Dataset #2 (04/2024)
    - T 2305
    - Q 587
    - R 16

Si osserva come la classe T sia quella maggiorente rappresentata con Q che comunque ha un numero di campioni decisamente inferiore. La classe come si vede è decisamente sottorappresentata in quanto è due ordini di grandezza inferiore rispetto alla classe "dominante" T.

2.3.3.3 YOLO object detector – fine tuning

ltralytics YOLOv8 è un modello all'avanguardia (SOTA) che si basa sul successo delle versioni precedenti di YOLO e introduce nuove funzionalità e miglioramenti per aumentare ulteriormente le prestazioni e la flessibilità. YOLOv8 è stato progettato per essere veloce, accurato e facile da usare, il che lo rende una scelta eccellente per un'ampia gamma di attività di rilevamento e tracciamento di oggetti, segmentazione di istanze, classificazione di immagini. Nel caso del presente lavoro di tesi si è partiti da una implementazione completa rilasciata da Ultralytics https://github.com/ultralytics/ultralytics.

Di seguito si riportano i risultati dalla fase di test dopo aver diviso il dataset usando la logica dell'80-10-10 ovvero 80% train, 10% validation e 10% test. Delle immagini di test sono state aumentante utilizzando la tecnica della data augmentation per aver un maggior numero di campioni anche considerando la dimensione del data-set. Di seguito si riportano gli iper-parametri usati.

- **imgsz (image size)**: 640
    - Dimensione delle immagini in pixel utilizzate
- **batch (batch size)**: 16
    - Numero di immagini per batch durante l'addestramento.

- **epochs**: 120
  - Numero di epoche di addestramento.
- **patience**: 50
  - Numero di epoche senza miglioramento prima di fermare l'addestramento anticipatamente.
- **optimizer**: 'SGD'
  - Algoritmo di ottimizzazione utilizzato (SGD - Stochastic Gradient Descent).
- **lr0 (initial learning rate)**: 0.01
  - Tasso di apprendimento iniziale.
- **lrf (final learning rate)**: 0.0001
  - Tasso di apprendimento finale.
- **momentum**: 0.937
  - Momento utilizzato per l'aggiornamento dei pesi.
- **weight_decay**: 0.0005
  - Decadimento dei pesi per la regolarizzazione L2.
- **warmup_epochs**: 3.0
  - Numero di epoche di riscaldamento con un tasso di apprendimento crescente.
- **warmup_momentum**: 0.8
  - Momento iniziale durante il riscaldamento.
- **warmup_bias_lr**: 0.1
  - Tasso di apprendimento del bias durante il riscaldamento.

2.3.4 *Risultati del modello addestrato.*

Di seguito si riportano i risultati del training i due differenti configurazioni:

- Data-set 1
- Data-set 1 + Data-set 2

Il primo data-set utilizzato ha mostrato una serie di limiti dovuti essenzialmente dall'utilizzo di un ridotto numero di immagini. Il secondo test è stato effettuato per valutare l'effetto di incrementare il numero di immagini e dunque di campioni al fine di incrementare i risultati circa alcune classi come ad esempio la classe Q.

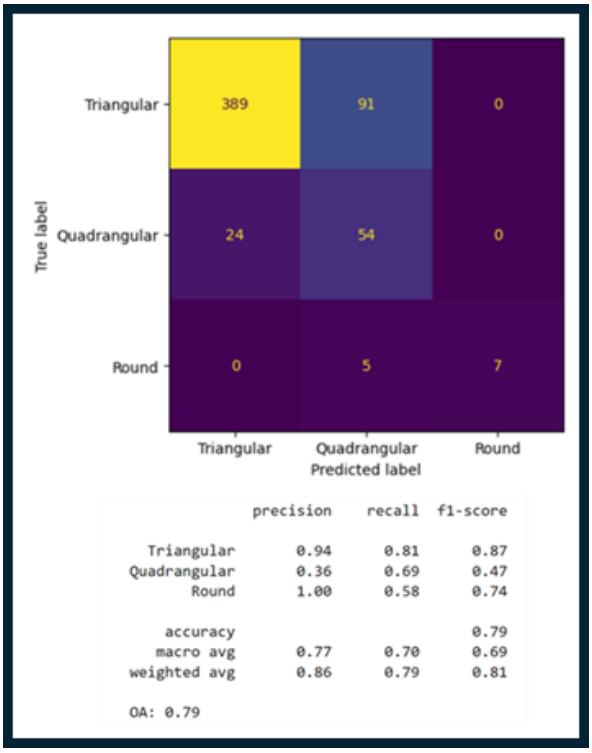
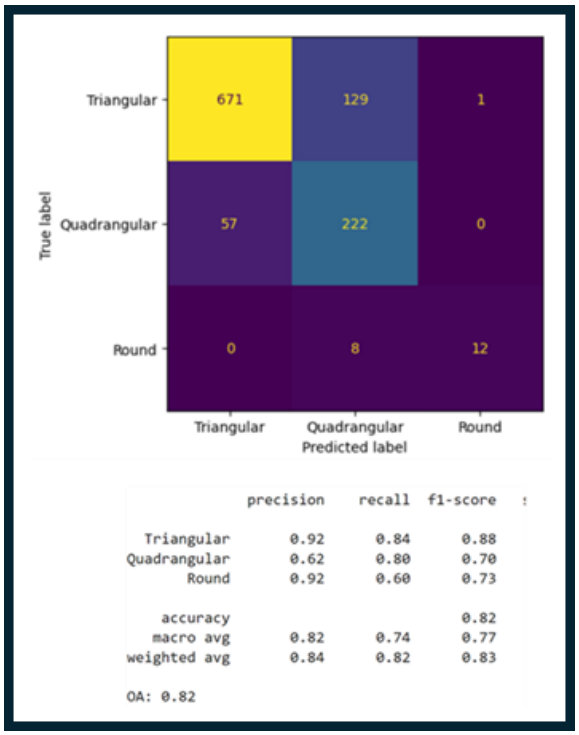


Figura 47: Matrici di confusione per il dataset iniziale.

Il secondo modello mostra come la dimensione superiore del dataset ha determinato un incremento di prestazione anche se rimane lo sbilanciamento tra classi dovuto ad una maggiore disponibilità di immagini annotate con foglie di tipo T. Seppur il miglioramento sia presente è necessario continuare ad alimentare il modello attraverso maggiori esempi di classi Q e R che sono comunque ancora sottorappresentate. Va sottolineato che il training del modello richiede molto tempo.

Ciò è stato un problema in quanto le risorse gratuite messe a disposizione da Google Colab hanno limiti sul numero di ore di uso continuative delle risorse “speciali” come ad esempio GPU (ndr. 12h). Si è risolto questo problema usando delle risorse messe a disposizione da UNIVPM sotto forma di GPU di tipo NVDIA RTX2080Ti.

Tale sistema di calcolo non ha limiti di utilizzo, ma va sottolineato che una singola sessione di training può richiedere fino a 96h di computazione.



	precision	recall	f1-score	
Triangular	0.92	0.84	0.88	
Quadrangular	0.62	0.80	0.70	
Round	0.92	0.60	0.73	
accuracy			0.82	
macro avg	0.82	0.74	0.77	
weighted avg	0.84	0.82	0.83	

OA: 0.82

Figura 48: Matrici di confusione per il dataset #1 e dataset #2.

I risultati ottenuti comunque dimostrano l’importanza di avere data-set che siano i più rappresentativi possibili delle classi “target”; è altresì evidente come lo sbilanciamento si un problema rilevante che apre comunque a scenari di altra natura, che fanno ad esempio riferimento a tecniche di tipo generativo come Generative Adversarial Networks (GANs).

CONCLUSIONI

L'identificazione di strategie efficaci per la conservazione, caratterizzazione e valorizzazione delle risorse genetiche agrarie è una sfida cruciale nell'ambito della ricerca scientifica legata al miglioramento genetico. La preservazione della biodiversità è infatti fondamentale anche per garantire uno sviluppo sostenibile ed una maggiore sicurezza alimentare, ovvero produzioni agricole adeguate a soddisfare il crescente bisogno di cibo di qualità a livello mondiale. Nel presente elaborato è stato analizzato un approccio innovativo basato sulla partecipazione dei cittadini (Citizen Science Experiment; CSE) al fine di studiare l'ampia biodiversità contenuta nelle risorse genetiche della specie *Phaseolus vulgaris.* Questo studio è parte di un progetto europeo ambizioso, denominato INCREASE (https://www.pulsesincrease.eu/) che mira ad identificare strategie innovative per preservare e valorizzare le risorse genetiche agrarie nelle leguminose alimentari. Tramite il CSE, i cittadini coinvolti a livello europeo nei primi tre round dell'esperimento hanno registrato un elevato numero di dati, relativi a caratteri agronomici e fenotipici. Nel presente elaborato è stata riportata un'analisi preliminare dei dati raccolti, che ha permesso di comprendere le dinamiche e le potenzialità di un approccio di conservazione decentralizzata delle risorse genetiche, utilizzando come specie modello, il fagiolo comune (*Phaseolus vulgaris*). Tramite il CSE, è stato possibile investigare sviluppo e adattamento di oltre 1000 accessioni di fagiolo comune in un'ampia varietà di territori all'interno dell'Unione Europea, grazie alla partecipazione dei cittadini; questo ha permesso di studiare la variabilità fenotipica dei materiali selezionati, in relazione a condizioni e fattori ambientali spesso molto differenti.

L'analisi dei dati raccolti dai cittadini, non comunemente coinvolti in processi di ricerca, ha mostrato in realtà il potenziale dell'impiego di approcci di scienza partecipata; infatti, grazie all'analisi preliminare condotta sui dati di fioritura disponibili dai primi tre round del CSE, i dati registrati dai cittadini replicano sostanzialmente scenari già noti in letteratura, sulla maggiore o minor precocità di alcune razze (identificabili anche come gruppi genetici distinti) entro la specie fagiolo comune. Infatti, nonostante i dati raccolti dai cittadini potrebbero non sempre essere rigorosi, la disponibilità di tutorial dettagliati e di un costante scambio di

informazioni tra scienziati e cittadini e tra cittadini, oltre alla grande mole di dati ottenuti, può garantire l'ottenimento di dati affidabili e replicabili.

Questo approccio partecipativo di conservazione decentralizzata delle risorse genetiche permette inoltre di raccogliere dati aggiuntivi relativi alla plasticità fenotipica (per quelle accessioni eventualmente replicate in più ambienti) e alla performance agronomica di specifici genotipi, che possono quindi risultare particolarmente interessanti per la coltivazione in areali specifici, o eventualmente per l'impiego in programmi di sviluppo varietale.

Nel presente elaborato si è inoltre discusso della difficoltà di analizzare e comparare dati provenienti da ambienti diversi (i.e., diverse migliaia di cittadini tra i vari round) con date di semina ovviamente non uniformi. A tal proposito, in aggiunta all'utilizzo di una varietà di controllo precoce distribuita tra tutti i cittadini, per la normalizzazione di dati fenologici (così come discusso nei risultati del presente elaborato), in un'ottica futura l'analisi dei dati potrebbe beneficiare sostanzialmente della scomposizione degli ambienti in variabili semplici. Ad esempio, si potrebbe ricorrere all'impiego di dati giornalieri per variabili ambientali (e.g., temperatura media, durata del giorno, quantità di pioggia) associati alle coordinate dei cittadini; in tal modo si potrebbero esprimere caratteri come la fioritura in funzione di fattori o indici correlati a tali variabili (e.g., Fioritura in gradi giorno [GDD; growing degree days], o in funzione delle ore di luce).

I dati ottenuti potranno poi supportare l'identificazione di marcatori molecolari per caratteri legati all'adattamento ed in funzione di variabili ambientali specifiche e correlate con tali caratteri, mediante analisi di associazione fenotipo-marcatore (GWAS; Genome Wide Association Study).

Il presente lavoro di tesi inoltre riporta l'impiego di una innovativa metodologia di fenotipizzazione supportata dall'intelligenza artificiale sulla base delle immagini raccolte dai singoli cittadini. Un modello di machine learning è stato infatti "allenato" grazie ad oltre 6000 immagini di foglie di fagiolo, al fine di renderlo capace di riconoscere il carattere "forma della foglia". In particolare, grazie alle fasi di training e validazione, il modello è già ad oggi in grado di assegnare correttamente, tra tre possibili forme (i.e., triangolare, rotonda e quadrangolare), l' 80% delle immagini di foglie disponibili in maniera corretta. Il modello potrà ovviamente beneficiare di un continuo allenamento, grazie anche alle immagini raccolte dai cittadini e disponibili dal quarto round in poi.

Il lavoro di analisi basato su intelligenza artificiale per l'identificazione di aspetti specifici come la forma delle foglie ha dimostrato l'enorme potenziale di queste tecnologie nel campo dell'elaborazione delle immagini. I risultati ottenuti evidenziano chiaramente l'importanza di

disporre di dataset che siano il più rappresentativi possibile delle classi "target". Un dataset ben bilanciato e ricco di variazioni è cruciale per l'addestramento efficace dei modelli di IA, poiché consente di migliorare l'accuratezza e la generalizzazione delle predizioni.

Tuttavia, è emerso che lo sbilanciamento dei dati (in termini di numero di istanze per una classe target) rappresenta un problema significativo. Dataset sbilanciati possono portare a modelli che sono incapaci di riconoscere correttamente tutte le classi di interesse, favorendo quelle più rappresentate. Questo problema non solo compromette l'efficacia del modello, ma limita anche la sua applicabilità in scenari reali.

Per affrontare queste sfide, le tecniche generative, come le Generative Adversarial Networks (GANs), offrono soluzioni promettenti. Le GANs possono essere utilizzate per generare dati sintetici che arricchiscono i dataset esistenti, riducendo così lo sbilanciamento e migliorando la rappresentatività delle classi minoritarie. Questo approccio non solo aumenta la quantità di dati disponibili per l'addestramento, ma può anche introdurre variazioni utili che rendono i modelli più robusti e versatili.

In sintesi, il lavoro svolto sottolinea l'importanza cruciale di dataset rappresentativi e bilanciati per il successo dei modelli di intelligenza artificiale nell'analisi delle immagini. Al contempo, evidenzia come le tecniche generative, come le GANs, possano offrire soluzioni innovative per superare le limitazioni dei dataset sbilanciati, aprendo la strada a ulteriori sviluppi e applicazioni nel campo dell'IA. Questi avanzamenti contribuiranno a migliorare la precisione e l'affidabilità dei sistemi di riconoscimento in contesti di fenotipizzazione anche in contesti di “Citizen Science” (ambienti di acquisizione delle immagini meno controllati rispetto ad un ambiente di laboratorio con esperti di settore).

In conclusione, il presente elaborato intende sottolineare l’importanza e le potenzialità degli approcci partecipativi nella ricerca scientifica, anche sulla base dei risultati preliminari riportati; infatti, grazie al coinvolgimento dei cittadini è stato innanzitutto possibile ottenere una quantità di dati affidabili, altrimenti non ottenibile mediante approcci classici, su un ampio set di risorse genetiche. Inoltre, mediante esperimenti come il CSE è possibile incrementare l’interesse e la consapevolezza verso l’importanza delle risorse genetiche, della biodiversità in generale, in specie che potrebbero ricoprire un ruolo ancora maggiore nell’alimentazione, con evidenti benefici. Il CSE rappresenta inoltre una valida strategia di conservazione, di tipo dinamico, che consente di valorizzare il potenziale genetico racchiuso nelle banche del germoplasma, con la possibilità per i cittadini di interagire, scambiare materiali e quindi diffondere accessioni particolarmente interessanti. È bene infine sottolineare come il cittadino sia il consumatore finale; conoscerne i gusti, le preferenze, stabilendo una diretta interazione

e promuovendo una partecipazione attiva in processi di conservazione e caratterizzazione delle risorse genetiche è quindi un valore aggiunto che può dare ulteriore impulso alla valorizzazione delle risorse genetiche agrarie.

BIBLIOGRAFIA

-Alesi Mattia; “Ottimizzazione di un algoritmo per una edge TPU al fine di riconoscere il grado di stanchezza di un guidatore mediante immagini RGB” 2022

-Alexander,D.H.,Novembre,J.&Lange,K.Fastmodel-basedesti mation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664 (2009).

-Alif, M. A. R., & Hussain, M. (2024). YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain. arXiv preprint arXiv:2406.10139.

-Al-Samarai , F. R. and Al-Kazaz, A. A. (2015). Molecular Markers: an Introduction and Applications. European Journal of Molecular Biotechnology, e002. doi: http://dx.doi.org/10.13187/ejmb.2015.9.118.

-Ariani A, Berny Mier Y Teran JC, Gepts P. Spatial and Temporal Scales of Range Expansion in Wild Phaseolus vulgaris. Mol Biol Evol. 2018 Jan 1;35(1):119-131. doi: 10.1093/molbev/msx273. PMID: 29069389; PMCID: PMC5850745.

-Ashby, J. A. (2009). The impact of participatory plant breeding. Plant breeding and farmer participation, 649-671.

-Atchison, G. W., Nevado, B., Eastwood, R. J., Contreras-Ortiz, N., Reynel, C., Madriñán, S., ... & Hughes, C. E. (2016). Lost crops of the Incas: Origins of domestication of the Andean pulse crop tarwi, Lupinus mutabilis. American Journal of Botany, 103(9), 1592-1606.

-Bellucci, E., Bitocchi, E., Rau, D., Rodriguez, M., Biagetti, E., Giardini, A., ... & Papa, R. (2014). Genomics of origin, domestication and evolution of Phaseolus vulgaris. Genomics of Plant Genetic Resources: Volume 1. Managing, sequencing and mining genetic resources, 483-507.

-Bellucci, E., Mario Aguilar, O., Alseekh, S., Bett, K., Brezeanu, C., Cook, D., De la Rosa, L., Delledonne, M., Dostatny, D.F., Ferreira, J.J., Geffroy, V., Ghitarrini, S., Kroc, M., Kumar Agrawal, S., Logozzo, G., Marino, M., Mary-Huard, T., McClean, P., Meglič, V., Messer, T., Muel, F., Nanni, L., Neumann, K., Servalli, F., Străjeru, S., Varshney, R.K., Vasconcelos, M.W., Zaccardelli, M., Zavarzin, A., Bitocchi, E., Frontoni, E., Fernie, A.R., Gioia, T., Graner, A., Guasch, L., Prochnow, L., Oppermann, M., Susek, K., Tenaillon, M. and Papa, R. (2021),

The INCREASE project: Intelligent Collections of food-legume genetic resources for European agrofood systems. Plant J, 108: 646-660. https://doi.org/10.1111/tpj.15472

-Bellucci, E., Benazzo, A., Xu, C. et al. Selection and adaptive introgression guided the complex evolutionary history of the European common bean. Nat Commun 14, 1908 (2023). https://doi.org/10.1038/s41467-023-37332-z

-Bhandari HR, Bhanu AN, Srivastava K, et al. Assessment of genetic diversity in crop plants - an overview. Adv Plants Agric Res. 2017;7(3):279-286. DOI: 10.15406/apar.2017.07.00255

-Biodiversity International 2009

-Bitocchi, E. (2004): Struttura della diversità genetica di varietà locali di pomodoro (Lycopersicon esculentum Mill). Tesi di Laurea, Università Politecnica delle Marche, Facoltà di Agraria, Corso di Laurea in Scienze e Tecnologie Agrarie, Dipartimento di Biotecnologie Agrarie ed Ambientali.

-Bitocchi E, Bellucci E, Giardini A, Rau D, Rodriguez M, Biagetti E, Santilocchi R, Spagnoletti Zeuli P, Gioia T, Logozzo G, Attene G, Nanni L, Papa R. Molecular analysis of the parallel domestication of the common bean (Phaseolus vulgaris) in Mesoamerica and the Andes. New Phytol. 2013 Jan;197(1):300-313. doi: 10.1111/j.1469-8137.2012.04377.x. Epub 2012 Nov 5. PMID: 23126683.

-Bitocchi E, Nanni L, Bellucci E, Rossi M, Giardini A, Zeuli PS, Logozzo G, Stougaard J, McClean P, Attene G, Papa R. Mesoamerican origin of the common bean (Phaseolus vulgaris L.) is revealed by sequence data. Proc Natl Acad Sci U S A. 2012 Apr 3;109(14):E788-96. doi: 10.1073/pnas.1108973109. Epub 2012 Mar 5. PMID: 22393017; PMCID: PMC3325731.

- Bitocchi, E., Rau, D., Bellucci, E., Rodriguez, M., Murgia, M. L., Gioia, T., ... & Papa, R. (2017). Beans (Phaseolus ssp.) as a model for understanding crop evolution. Frontiers in plant science, 8, 251783.

-Bonney R. 1996. Citizen science: A lab tradition. Living Bird 15: 7–15

-Bonney R, Ballard HL, Jordan R, McCallie E, Phillips T, Shirk J and Wilderman CC (2009) Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. Washington, D.: Center for Advancement of Informal Science Education (CAISE).

-Bonney, R., Phillips, T. B., Ballard, H. L., & Enck, J. W. (2016). Can citizen science enhance public understanding of science?. Public understanding of science, 25(1), 2-16.

-Brown, Anthony. (1989). Core collections: A practical approach to genetic resource management. Genome. 31. 818-824. 10.1139/g89-144.

-Brown, G., & Yule, G. (1983). Discourse analysis. Cambridge: Cambridge University Press.

-Casavecchia Simona 2023; Gestione e tutela della biodiversità e del paesaggio

-Ceccarelli, S. (2006). Decentralized-participatory plant breeding: Lessons from the south-Perspectives in the north. Participatory Plant Breeding: Relevance for Organic Agriculture?, 8.

-Ceccarelli, S., & Grando, S. (2007). Decentralized-participatory plant breeding: an example of demand driven research. Euphytica, 155, 349-360.

-Cohn, J.P., 2008. Citizen science: can volunteers do real research? Bioscience 58, 192–197

-Collotta Martina 2018 Rivista Società Italiana di Medicina Generale

-Consiglio nazionale delle ricerche; CNR

-Convention on Biological Diversity; CBD. 1992

-Cortesi, F. (1933): Leguminose. Enciclopedia Italiana.

-Cortinovis Gaia, Valerio Di Vittori, Elisa Bellucci, Elena Bitocchi, Roberto Papa,

Adaptation to novel environments during crop diversification, Current Opinion in Plant Biology, Volume 56, 2020, Pages 203-217, ISSN 1369-5266, https://doi.org/10.1016/j.pbi.2019.12.011.

-Cortinovis, G., Oppermann, M., Neumann, K., Graner, A., Gioia, T., Marsella, M., ... & Bitocchi, E. (2021). Towards the development, maintenance, and standardized phenotypic characterization of single-seed-descent genetic resources for common bean. Current Protocols, 1(5), e133.

- Cortinovis, L. Vincenzi, R. Anderson, G. Marturano, J.I. Marsh, P.E. Bayer, L. Rocchet ti, G. Frascarelli, G. Lanzavecchia, A. Pieri, A Benazzo, E. Bellucci, V. DiVittori, L. Nanni, J.J. FerreiraFernández, M. Rossato, O.M. Aguilar, P.L. Morrell, M. Rodriguez, T. Gioia, K. Neumann, J.C. AlvarezDiaz, A. GratiasWeill, C. Klopp, V. Geffroy, E. Bitocchi, M. Delledonne, D. Edwards, R Papa.bioRxiv 2023.11.23.568464; doi: https://doi.org/10.1101/2 023.11.23.568464

-Desiderio, F., Bitocchi, E., Bellucci, E., Rau, D., Rodriguez, M., Attene, G., ... & Nanni, L. (2013). Chloroplast microsatellite diversity in Phaseolus vulgaris. Frontiers in Plant Science, 3, 39686.

-Dong, Z., Chen, Y. Transcriptomics: Advances and approaches. Sci. China Life Sci. 56, 960–967 (2013). https://doi.org/10.1007/s11427-013-4557-2

-Engels, J. M. M., & Maggioni, L. (2012). 41 AEGIS: A Regionally Based Approach to PGR Conservation. Agrobiodiversity conservation: securing the diversity of crop wild relatives and landraces, 321.

-Food and Agriculture organization of the United Nations FAO; Glossary https://www.fao.org/pgrfa/resources/documentlogs/Glossary_EN.pdf

-Food and Agriculture organization of the United Nations FAO; Summary of Food and Agricultural Statics 2004.

-Food and Agriculture organization of the United Nations FAO; FAO – WIEWS 2009.

-Freyre R, Ríos R, Guzmán L, Debouck D, Gepts P (1996) Ecogeographic distribution of Phaseolus spp. (Fabaceae) in Bolivia. Econ Bot 50:195–215

-Gambus, P., & Shafer, S. L. (2018). Artificial intelligence for everyone. Anesthesiology, 128(3), 431-433.

-Gao, L., Gonda, I., Sun, H. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 51, 1044–1051 (2019). https://doi.org/10.1038/s41588-019-0410-2

-Gepts, P., & Bliss, F. A. (1988). Dissemination pathways of common bean (Phaseolus vulgaris, Fabaceae) deduced from phaseolin electrophoretic variability. II. Europe and Africa. Economic Botany, 42(1), 86-104.

- Gepts P, Papa R, Coulibaly S, González Mejía A, Pasquet R (1999) Wild legumediversity and domestication—insights from molecular methods. Wild Legumes:Proceedings of the Seventh MAFF International Workshop on Genetic Resources, edVaughan D (National Institute of Agrobiological Resources, Tsukuba, Japan), pp19–31.

-Gladstones, J. S. (1998). Distribution, origin, taxonomy, history and importance. Lupins as crop plants: biology, production and utilization., 1-37.

-Gu R, Fan S, Wei S, Li J, Zheng S, Liu G. Developments on Core Collections of Plant Genetic Resources: Do We Know Enough? Forests. 2023; 14(5):926. https://doi.org/10.3390/f14050926

-Jose, D.M.; Lucía, D.L.R.; Isaura, M.; Luís, G.; Elena, C.M.; Cristina, M.; Joan, C.; Joan, S.; Ana, R.; German, A.; et al. Plant genebanks: Present situation and proposals for their improvement. the case of the Spanish network. Front. Plant Sci. 2018, 9, 1794.

-Joshi Bal, Ghimire Krishna, Shrestha Deepa - 2018/12/17 SN - 978-9937-0-5413-3. The National Genebank's Promotion of Community Seed Banks: Status and Strategy

-Harlan, J. R. (1992). Crops and man.

-Heywood, V.H. (1995) Global Biodiversity Assessment. United Nations Environment Programme. Cambridge University Press, Cambridge.

-Heuermann, M.C., Knoch, D., Junker, A. et al. Natural plant growth and development achieved in the IPK PhenoSphere by dynamic environment simulation. Nat Commun 14, 5783 (2023). https://doi.org/10.1038/s41467-023-41332-4

-Hintum, T. V., & Knüpffer, H. (1995). Duplication within and between germplasm collections. I. Identifying duplication on the basis of passport data.

-Intelligent Collections of Food Legumes Genetic Resources for European Agrofood Systems (INCREASE). https://www.pulsesincrease.eu/

-International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005 Aug 11;436(7052):793-800. doi: 10.1038/nature03895. PMID: 16100779.

-Irwin, A. (1995) Citizen Science: A Study of People, Expertise and Sustainable Development. Routledge, London.

-Istituto Superiore per la protezione e la ricerca ambientale; ISPRA; https://www.isprambiente.gov.it/it

-Jasanoff S (2004) States of Knowledge: The Co-Production of Science and Social Order. London and New York: Routledge.

- Kami J, Velásquez VB, Debouck DG, Gepts P (1995) Identification of presumedancestral DNA sequences of phaseolin in Phaseolus vulgaris. Proc Natl Acad Sci USA92:1101–1104.

-Leibniz Institute of Plant Genetics and Crop Research (IPK). https://www.ipk-gatersleben.de/en/

-Li L, Zhang Q, Huang D. A Review of Imaging Techniques for Plant Phenotyping. Sensors. 2014; 14(11):20078-20111. https://doi.org/10.3390/s141120078

-Lioi Lucia, Piergiovanni R. Angela (2013); CNR, Istituto di Genetica Vegetale, Bari, Italia

-Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z. Pan-Genome of Wild and Cultivated Soybeans. Cell. 2020 Jul 9;182(1):162-176.e13. doi: 10.1016/j.cell.2020.05.023. Epub 2020 Jun 17. PMID: 32553274.

-Lorenzetti F. et al., 2021; (Miglioramento genetico delle piante) Edagricole

-Magrini Marie-Benoit , Anton Marc , Chardigny Jean-Michel , Duc Gerard , Duru Michel , Jeuffroy Marie-Helene , Meynard Jean-Marc , Micard Valerie , Walrand Stephane, Pulses for Sustainability: Breaking Agriculture and Food Sectors Out of Lock-In, Frontiers in Sustainable Food Systems VOLUME 2 (2018)

URL=https://www.frontiersin.org/journals/sustainable-foodsystems/articles/10.3389/fsufs.2018.0006410.3389/fsufs.2018.00064
ISSN=2571-581X

-Maphosa Y., Jideani V. A. (2017). The Role of Legumes in Human Nutrition, Functional Fo–d - Improve Health through Adequate Food. Edited by Maria Chavarri Hueda, IntechOpen, DOI: 10.5772/intechopen.69127.

-Milner, S.G., Jost, M., Taketa, S. et al. Genebank genomics highlights the diversity of a global barley collection. Nat Genet 51, 319–326 (2019). https://doi.org/10.1038/s41588-018-0266-x

-Ministero dell'agricoltura, della sovranità alimentare e delle foreste; Masaf

-Mishra, K.B., Mishra, A., Klem, K. et al. Plant phenotyping: a perspective. Ind J Plant Physiol. 21, 514–527 (2016). https://doi.org/10.1007/s40502-016-0271-y

- Nabwire, S., Suh, H. K., Kim, M. S., Baek, I., & Cho, B. K. (2021). Application of artificial intelligence in phenomics. Sensors, 21(13), 4363.

-Newman Greg, Jim Graham, Alycia Crall, Melinda Laituri, The art and science of multi-scale citizen science support, Ecological Informatics, Volume 6, Issues 3–4, 2011, Pages 217-227, ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2011.03.002. (https://www.sciencedirect.com/science/article/pii/S157495411100029X)

-Nonic, M., Šijačić-Nikolić, M. (2021). Genetic Diversity: Sources, Threats, and Conservation. In: Leal Filho, W., Azul, A.M., Brandli, L., Lange Salvia, A., Wall, T. (eds) Life on Land. Encyclopedia of the UN Sustainable Development Goals. Springer, Cham. https://doi.org/10.1007/978-3-319-95981-8_53

-Pandotra, P.; Gupta, S. Biotechnological Approaches for Conservation of Plant Genetic Resources and Traditional Knowledge. In Plant Genetic Resources and Traditional Knowledge for Food Security; Salgotra, R.K., Gupta, B.B., Eds.; Springer: Berlin/Heidelberg, Germany, 2015.

-Paolinelli Giacomo (2022) Le basi genetiche della deiscenza del baccello in Phaseolus vulgaris

-Papa, R., Bellucci, E., Rossi, M., Leonardi, S., Rau, D., Gepts, P., ... & Attene, G. (2007). Tagging the signatures of domestication in common bean (Phaseolus vulgaris) by means of pooled DNA samples. Annals of Botany, 100(5), 1039-1051.

-Patriarca, E. J., Tatè, R., & Iaccarino, M. (2002). Key role of bacterial NH4+ metabolism in Rhizobium-plant symbiosis. Microbiology and Molecular Biology Reviews, 66(2), 203-222.

-Patti GJ, Yanes O, Shriver LP, Courade JP, Tautenhahn R, Manchester M, Siuzdak G. Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin. Nat Chem Biol. 2012 Jan 22;8(3):232-4. doi: 10.1038/nchembio.767. PMID: 22267119; PMCID: PMC3567618.

-Rajasekharan, P.E.; Sahijram, L. In vitro conservation of plant cermplasm. In Plant Biology and Biotechnology; Bahadur, B., Venkat Rajam, M., Sahijram, L., Krishnamurthy, K.V., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; Volume II: Plant Genomics and Biotechnology.

-Rao, V. R., & Riley, K. W. (1994). The use of biotechnology for conservation and utilization of plant genetic resources. Plant Genet. Res. Newsletter, 97(3).

-Redmon (2016) lookonceunifiedrealtime, title={You Only Look Once: Unified, Real-Time Object Detection}, author={Joseph Redmon and Santosh Divvala and Ross Girshick and Ali Farhadi}, year={2016}, eprint={1506.02640},archivePrefix={arXiv},primaryClass={cs.CV}, url={https://arxiv.org/abs/1506.02640},

-Rendón-Anaya , M., Montero-Vargas, J. M., Saburido-Álvarez, S., Vlasova, A., Capella-Gutierrez, S., Ordaz-Ortiz, J. J., ... & Herrera-Estrella, A. (2017). Genomic history of the origin and domestication of common bean unveils its closest sister species. Genome biology, 18, 1-17.

-Rocchetti, L., Gioia, T., Logozzo, G., Brezeanu, C., Pereira, L. G., la Rosa, L. D., ... & Papa, R. (2022). Towards the development, maintenance and standardized phenotypic characterization of single-seed-descent genetic resources for chickpea. Current Protocols, 2(2), e371.

-Roque, A., Wutich, A., Quimby, B., Porter, S., Zheng, M., Hossain, M. J., & Brewis, A. (2022). Participatory approaches in water research: A review. Wiley Interdisciplinary Reviews: Water, 9(2), e1577.

-Salgotra RK, Chauhan BS. Genetic Diversity, Conservation, and Utilization of Plant Genetic Resources. Genes. 2023; 14(1):174. https://doi.org/10.3390/genes14010174

-Santalla, M., Rodiño, A., & De Ron, A. (2002). Allozyme evidence supporting southwestern Europe as a secondary center of genetic diversity for the common bean. Theoretical and applied genetics, 104, 934-944.

-Scheben A, Batley J, Edwards D. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J. 2017 Feb;15(2):149-161. doi: 10.1111/pbi.12645. PMID: 27696619; PMCID: PMC5258866.

-Schmutz, J., McClean, P., Mamidi, S. et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46, 707–713 (2014). https://doi.org/10.1038/ng.3008

-Shapin S and Schaffer S (1985) Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life. Princeton: Princeton University Press.

-Simmonds, N. W. (1991). Selection for local adaptation in a plant breeding programme. Theoretical and Applied Genetics, 82, 363-367.

-Singh, S. P., Gepts, P., & Debouck, D. G. (1991). Races of common bean (Phaseolus vulgaris, Fabaceae). Economic Botany, 379-396.

-Singh Arti, Baskar Ganapathysubramanian, Asheesh Kumar Singh, Soumik Sarkar,

Machine Learning for High-Throughput Stress Phenotyping in Plants, Trends in Plant Science, Volume 21, Issue 2, 2016, Pages 110-124, ISSN 1360-1385, https://doi.org/10.1016/j.tplants.2015.10.015.

-Sivasubramanian M.; Artificial Intelligemce's impact on our everyday lives (2021). "Learning outcomes of classrom research" book

-Song, JM., Guan, Z., Hu, J. et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat. Plants 6, 34–45 (2020). https://doi.org/10.1038/s41477-019-0577-7

-Stevens, P. F. (2017). Angiosperm Phylogeny Website. Version 14, July 2017.

-Strasser, B. J., Baudry, J., Mahr, D., Sanchez, G. and Tancoigne, E. (2019) "'Citizen Science'? Rethinking Science and Public Participation", Science & Technology Studies, 32(2), pp. 52–76. doi: 10.23987/sts.60425.

-Swingland Ian R., Biodiversity, Definition of, Editor(s): Simon A Levin, Encyclopedia of Biodiversity (Second Edition), Academic Press, 2013, Pages 399-410, ISBN 9780123847201, https://doi.org/10.1016/B978-0-12-384719-5.00009-5

- Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012 May 30;485(7400):635-41. doi: 10.1038/nature11119. PMID: 22660326; PMCID: PMC3378239.

-Torkamaneh, D., Boyle, B. & Belzile, F. Efficient genome-wide genotyping strategies and data integration in crop plants. Theor Appl Genet 131, 499–511 (2018). https://doi.org/10.1007/s00122-018-3056-z

-Toro, O., Tohme, J., & Debouck, D. (1990). Wild bean (Phaseolus vulgaris L.): description and distribution (Vol. 181). CIAT.

-Trachsel, S., Kaeppler, S.M., Brown, K.M. et al. Shovelomics: high throughput phenotyping of maize (Zea mays L.) root architecture in the field. Plant Soil 341, 75–87 (2011). https://doi.org/10.1007/s11104-010-0623-8

-Van Etten j, Beza E, Calderer L, et al. first experiences with a novel farmer citizen science approach: crowdsourcing participatory variety selection through on-farm triadic comparisons of technologies (tricot). experimental agriculture. 2019;55(s1):275-296. doi:10.1017/s0014479716000739

-Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., ... & Cook, D. R. (2013). Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nature biotechnology, 31(3), 240-246.

-Vavilov, N. I. (1926). Studies on the Origin of Cultivated Plants... Institut de Botanique Appliquée et d'Amélioration des Plantes.

-Verma, M., Kulshrestha, S., Puri, A. (2017). Genome Sequencing. In: Keith, J. (eds) Bioinformatics. Methods in Molecular Biology, vol 1525. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-6622-6_1

-Wambugu, P. W., Ndjiondjop, M. N., & Henry, R. J. (2018). Role of genomics in promoting the utilization of plant genetic resources in genebanks. Briefings in functional genomics, 17(3), 198-206.

-Wambugu PW, Nyamongo DO, Kirwa EC. Role of Seed Banks in Supporting Ecosystem and Biodiversity Conservation and Restoration. Diversity. 2023; 15(8):896. https://doi.org/10.3390/d15080896

-Weise Stephan, Markus Oppermann, Lorenzo Maggioni, Theo van Hintum, Helmut Knüpffer, EURISCO: The European search catalogue for plant genetic resources, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D1003–D1008, https://doi.org/10.1093/nar/gkw755

-Whittaker R. H. 1972 Paper for "Origin and Measurement of Diversity," Summer Institute in Systematics V, Smithsonian Institution, Washington, D.C., 1971.

-Wilson A.G. et al. 1988; Applied Research and Development https://doi.org/10.1068/a200849

-Würschum, T., Leiser, W. L., Jähne, F., Bachteler, K., Miersch, M., & Hahn, V. (2019). The soybean experiment '1000 Gardens': a case study of citizen science for research, education, and beyond. Theoretical and Applied Genetics, 132, 617-626.

-Xu Z, York LM, Seethepalli A, Bucciarelli B, Cheng H, Samac DA. Objective Phenotyping of Root System Architecture Using Image Augmentation and Machine Learning

in Alfalfa (Medicago sativa L.). Plant Phenomics. 2022 Apr 7;2022:9879610. doi: 10.34133/2022/9879610. PMID: 35479182; PMCID: PMC9012978.

-Yang, D., Du, X., Yang, Z., Liang, Z., Guo, Z. and Liu, Y. (2014), Transcriptomics, proteomics, and metabolomics to reveal mechanisms underlying plant secondary metabolism. Eng. Life Sci., 14: 456-466. https://doi.org/10.1002/elsc.201300075

-Zhu Jinming, Paul A Ingram, Philip N Benfey, Tedd Elich, From lab to field, new approaches to phenotyping root system architecture, Current Opinion in Plant Biology, Volume 14, Issue 3, 2011, Pages 310-317, ISSN 1369-5266, https://doi.org/10.1016/j.pbi.2011.03.020.

-Zohary, D., Hopf, M., & Weiss, E. (2012). Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin. Oxford University Press.