CONVOCATORIA 811 (PROGRAMA DE ESTANCIAS POSTDOCTORALES BENEFICIARIOS DE FORMACIÓN COLCIENCIAS 2018 EN ENTIDADES DEL SNCTeI)

PROPUESTA DE INVESTIGACIÓN

En este documento se presenta la propuesta de investigación estructurada en la que se describen los términos y las condiciones en los cuales se desarrollará la estancia posdoctoral. Se destaca el aporte que hará el doctor y cómo su contribución se traduciría en una mejora dentro del proyecto.

Resumen

La selección genómica (SG) es una estrategia para acelerar el mejoramiento de caracteres cuantitativos poligénicos. Ha sido ampliamente usada en ganadería y más recientemente en cultivos con excelentes resultados. El objetivo de la SG es crear modelos para analizar y asociar una enorme cantidad de información genómica y fenotípica con el fin de predecir el desempeño de características de interés agronómico en las progenies ya sea de poblaciones de mejoramiento genético o de los bancos de germoplasma. En Colombia, esta herramienta está siendo implementada con mucho éxito en ganadería, pero los cultivos más importantes del país, incluyendo la papa, todavía está en desarrollo. AGROSAVIA tiene a su cargo la Colección Central Colombiana de papa (CCC). A través de muchos años de investigación, las 800 accesiones que hacen parte de la CCC cuentan hoy en día con una excelente base de datos de caracterización genómica y fenotípica. Por esta razón, la CCC es un modelo excelente para desarrollar un programa de SG enfocado en bancos de germoplasma. Para este proyecto, nos planteamos como pregunta de investigación ¿Cuál es la estrategia computacional óptima para hacer predicciones de características de interés agronómico en papas nativas diploides y tetraploides, utilizando información genómica? Al final del proyecto, esperamos encontrar el modelo más óptimo y eficiente para (1) seleccionar en la CCC características morfológicas de interés agroindustrial en papa y (2) encontrar resistencia horizontal a gota y polilla guatemalteca, dos de los patógenos que generan las pérdidas más altas en la producción de papa del país.

Abstract

The Genomic selection (GS) is one of the most recent strategies to accelerate the improvement of polygenic quantitative characters in breeding programs. It has been widely used in animal breeding and more recently in plant breeding with excellent results. The objective of the SG is to develop models to analyze and associate a massive amount of genomic and phenotypic information to predict the performance of characteristics of agronomic interest in the progenies of either breeding populations or germplasm banks. In Colombia, the implementation of GS is very advanced in livestock programs, but for the most important crops in the country, including potatoes, are still under development. AGROSAVIA is in charge of the Central Colombian Potato Collection (CCC). Through many years of research, the 800 accessions CCC has an excellent database of genomic and phenotypic characterization. Therefore, the CCC is a unique model to develop a GS program focused on a national germplasm bank. Our primary objective for this project is to determine what is the optimal computational strategy to make predictions for important traits in both in diploid and tetraploid native potatoes using genomic information. At the end of the project, we hope to find the most optimal and efficient model to (1) select morphological characteristics of industrial interest in potatoes and to (2) find resistance to Guatemalan drop and moth, two of the pathogens that generate the highest losses in the country's potato production.

Título de la propuesta
Evaluación de estrategias de selección genómica y GWAS en papas nativas diploides y tetraploides conservadas en la Colección Central Colombiana de papa (CCC)

Área de la propuesta: Bioinformática

Gran área de conocimiento	1 Ciencias Naturales
Área de conocimiento	1.B Computación y ciencias de la información
Disciplina de conocimiento	1B02 Ciencias de la información y bioinformática

Nombre de la entidad
Corporación Colombiana de Investigación Agropecuaria – AGROSAVIA
Contacto: María Aidé Londoño Arias. Correo: alondono@agrosavia.co. Teléfono: (+57 1) 4227300, extensión 1265, 1574, 1575.

Nombre del doctor o de los doctores que va(n) a desarrollar su estancia posdoctoral	**Tipo y número de identificación**
Luis Ernesto Garreta Unigarro	C.C. 87712548

Nombre del Investigador de Corpoica responsable de la propuesta	**Tipo y número de identificación**
Paula Helena Reyes Herrera Ivania Cerón Souza	C.C. 52992302 C.C. 59822714

Tema de investigación
Uso de la bioinformática para acelerar programas de mejoramiento de papas nativas

Objetivo general
Identificar una estrategia óptima para asociar SNPs y características de interés agronómico en papas nativas diploides y tetraploides utilizando algoritmos de selección genómica (SG) y GWAS (Genome-Wide Assisted Selection).

Objetivos específicos

- Comparar cuantivamente el desempeño de algoritmos de selección genómica utilizando tres ramas de algoritmos aplicados actualmente (BLUP, Bayesianas y Machine Learning).
- Comparar cuantitativamente el desempeño de los algoritmos de selección genómica y GWAS para características gobernadas por muchos genes vs. pocos genes.
- Identificar si existe una estrategia de selección genómica que se pueda aplicar a papas diploides y tetraploides simultáneamente.

Justificación

El gobierno colombiano a través del COMPES 3850, propone promover una visión integral del territorio, en donde se incluye la conservación y el aprovechamiento sostenible de la diversidad del país. Dentro de este gran objetivo nacional, uno de los retos más importantes es la caracterización de la agrobiodiversidad, como un componente importante de la biodiversidad y así mismo, con base en este conocimiento, la utilización de herramientas de innovación que nos permitan un mejor aprovechamiento de los recursos naturales y la producción rural, especialmente en zonas de conflicto armado[1].

La papa es uno de los cultivos más importantes del país. Para 2016, la papa en Colombia representó 149.744 ha de área sembrada/área plantada, con una producción anual de 3.034.031 toneladas[2]. Adicionalmente, esta especie hace parte de la dieta básica colombiana con alto potencial para mejorar la seguridad alimentaria del país y la economía campesina[3]. Sin embargo, la producción de papa del país sufre de diversos problemas fitosanitarios, los cuales son de gran interés en investigación porque pueden generar pérdidas totales en la cosecha. En este sentido, se ha hecho un gran esfuerzo en investigación a nivel nacional e internacional para identificar cuáles son las bases genéticas cuantitativas de la resistencia a varias enfermedades que se agravan con el cambio climático como gota y polilla guatemalteca[4,5], así como las bases genéticas de la gran diversidad de características morfológicas que se encuentran en las papas nativas de Colombia[6].

AGROSAVIA administra la Colección Central Colombiana de papa (CCC). Esta colección, que es considerada la segunda más biodiversa del mundo cuenta con más de 2.000 accesiones provenientes de diversas regiones del país. En el año 2017, AGROSAVIA publicó la caracterización molecular de la CCC utilizando más de 5.000 SNPs a lo largo del genoma. Mediante este análisis se corroboró la gran diversidad de la colección y así mismo, una densidad de SNP suficientes para hacer estudios de asociación fenotipo-genotipo[6]. Adicionalmente, la CCC cuenta con una excelente caracterización fenotípica en campo de caracteres morfológicos

y también asociada a los patógenos que atacan la papa como gota y polilla guatemalteca.

Usando como base todo ese conocimiento acumulado, la CCC es un excelente modelo para comenzar a utilizar metodologías y estrategias de la selección asistida por marcadores (SAM) como una manera de conocer y aprovechar la diversidad genética y fenotípica almacenada en los bancos de germoplasma de la nación colombiana. Específicamente, por medio de este proyecto proponemos la comparación cuantitativa del desempeño y de la exactitud de la predicción utilizando diferentes estrategias de Selección Genómica (SG) y de GWAS (Genome-Wide Assisted selection, por sus siglas en inglés). Con estos datos, esperamos identificar casos óptimos para modelos de SG y para Selección Asistida por Marcadores (SAM) en papa para realizar predicciones en la progenie sobre características morfológicas de interés agroindustrial y de resistencia a problemas fitosanitarios más limitantes en la producción de papa en Colombia como son gota y polilla guatemalteca[7].

Para este proyecto, AGROSAVIA cuenta con un grupo de investigación multidisciplinario adscrito a Colciencias titulado “Mejoramiento Genético Vegetal, Uso y Aprovechamiento de la Agrobiodiversidad (MGVA)” que le dará apoyo permanente al postdoc para culminar su proyecto con éxito. El postdoc contará con la asesoría de los siguientes investigadores PhD y Máster: Paula Reyes (Bioinformática), Ivania Cerón (Genética de Poblaciones), Roxana Yockteng (Biología Evolutiva de plantas), Jhon Berdugo (Genómica de plantas que actualmente están trabajando en la caracterización genómica de la CCC) y de los investigadores fitopatólogos que trabajan en papa Mauricio Soto, Liliana Cely y Juan David Santa, expertos en las enfermedades de gota y polilla guatemalteca.

Marco conceptual

La selección genómica (**SG** o GS, del inglés "Genomic Selection") es una estrategia para acelerar el mejoramiento de caracteres cuantitativos poligénicos. Ha sido ampliamente usada en animales y más recientemente en plantas[8]. La **SG** utiliza datos genotípicos y fenotípicos para calcular una medida cuantitativa del valor de cada genotipo como parental para los futuros ciclos de mejoramiento. Este valor se conoce como el valor de mejoramiento genómico estimado (**VMGE o "breeding value" en inglés**). En la práctica, la **SG** tiene una serie de pasos recurrentes. Primero, los datos genotípicos y fenotípicos son colectados de un conjunto de entrenamiento (**CE**), el cual está conformado por individuos que hacen parte de una población de referencia o población de prueba en la cual se realizará la selección. Posteriormente, estos datos son utilizados para construir modelos estadísticos que matemáticamente relacionan el genotipo con el fenotipo en el **CE**.

En el esquema de **SG**, una vez el modelo ha sido desarrollado, aquellos individuos con valores más altos de **VMGE** pueden ser seleccionados de la población de prueba a partir de los datos genotípicos únicamente. Por tanto, los datos genotípicos de cada individuo de la población de prueba son los datos de entrada para el modelo de **SG** que calculará los **VMGE**. Finalmente, los individuos con altos **VMGE** son seleccionados y usados como parentales y las progenies continuarán el proceso de mejoramiento y selección[9,10].

En la actualidad existen tres familias grandes de modelos estadísticos para hacer predicción y SG. Esas tres familias son (1) BLUP (Best Linear Unbiased Prediction, por sus siglas en inglés), (2) modelos bayesianos y (3) modelos de machine learning[11]. Estas tres familias de modelos tratan de resolver problemas estadísticos asociados a la dimensionalidad de los datos, así como también mejorar su poder predictivo. En el caso de los BLUP, son lineales y asumen un modelo infinitesimal. Es decir, asumen que hay una gran cantidad de genes asociados a una característica fenotípica, cada uno con efectos pequeños y dispersos a lo largo de los cromosomas. En contraste, los modelos bayesianos asumen que hay un efecto diferencial en los genes. Es decir, que hay pocos genes con efectos muy grandes y muchos genes con efectos pequeños. Por lo tanto, el modelo da pesos diferenciales a cada uno de los genes asociados a la característica fenotípica. Finalmente, los modelos de machine learning cuyo foco es clasificar individuos entre clases dependiendo de la característica y así maximizar la probabilidad de un individuo de pertenecer a una clase [11].

En el caso de este proyecto, esperamos que el postdoc compare cuantitativamente el desempeño y la exactitud de la predicción de cada una de las familias de modelos predictivos con el fin de escoger cuál o cualés modelos se ajustan mejor para predecir características de interés agronómico de las accesiones de la CCC asociadas a la diversidad morfológica y a la resistencia a patógenos. Así mismo, esperamos comparar estos modelos predictivos con herramientas de GWAS con el objetivo de implementar dos herramientas complementarias en papa para hacer SAM de una manera eficiente y rápida[12].

Metodología de investigación

En este estudio, primero nos vamos a concentrar en seleccionar y ajustar el método óptimo para predecir VMGE a partir de marcadores moleculares. Para cumplir este objetivo se buscará seleccionar el método óptimo de predicción entre las familias de SG y GWAS siguiendo el diseño experimental de la Fig.1 que nos servirán para cumplir los primeros dos objetivos planteados en este proyecto.

Inicialmente vamos a aprovechar toda la información fenotípica asociada a la resistencia que muestran en campo las accesiones de la CCC hacia enfermedades como gota y polilla guatemalteca. Esta información fenotípica se va a combinar con información genómica generada sobre la CCC en uno de los proyectos de AGROSAVIA[6] (Fig. 1) que busca identificar individuos resistentes a polilla guatemalteca y a gota. Posteriormente nos vamos a extender a características morfológicas que han sido evaluadas en el BGV.

Para construir los modelos de análisis, nuestro **CE** corresponderá a todas las accesiones de la CCC, las cuales tienen información fenotípica y genotípica (Fig. 1). Sobre este conjunto de datos se hará la selección del método de predicción usando validación cruzada. Con esa información, se hará la predicción genómica usando los datos genotípicos del conjunto de validación (cruces entre papas nativas que en estudios previos han sido identificadas como resistentes a gota y a polilla guatemalteca). Combinando esta información de predicción genómica y validación, se obtendrá un modelo para la selección genómica de papa (Fig. 1).

Metodología para objetivo 1. Mediremos el desempeño de cada método evaluando la sensibilidad y especificidad que se obtiene tanto en el conjunto de entrenamiento como en el conjunto de validación. Para esto se ajustarán parámetros internos de cada método buscando comparar la configuración óptima usando validación cruzada en el conjunto de entrenamiento. Adicionalmente se tendrá en cuenta el esfuerzo computacional necesario para realizar la predicción como un criterio adicional. Buscaremos identificar ventajas y desventajas de cada método, y las condiciones que favorecen el uso de los métodos estudiados.

Metodología para objetivo 2. Identificaremos para cada característica el conjunto de marcadores con mayores efectos asociados utilizando GWAS. Adicionalmente, entrenaremos modelos de predicción genómica variando el número de marcadores considerados (ej. 100, 250, 500, 1000, 2000, 4000 y todos); en cada uno de los casos los marcadores serán seleccionados al azar y para cada número de marcadores se entrenarán al menos 10 modelos. Posteriormente compararemos en el conjunto de validación las inferencias que se pueden hacer (1) teniendo en cuenta los genes con mayores efectos asociados mediante GWAS y (2) la predicción genómica con modelos entrenados con diferentes números de marcadores. Buscaremos identificar para características gobernadas por pocos y por muchos genes que estrategia conviene usar y el número de marcadores en que comienza a existir alguna diferencia.

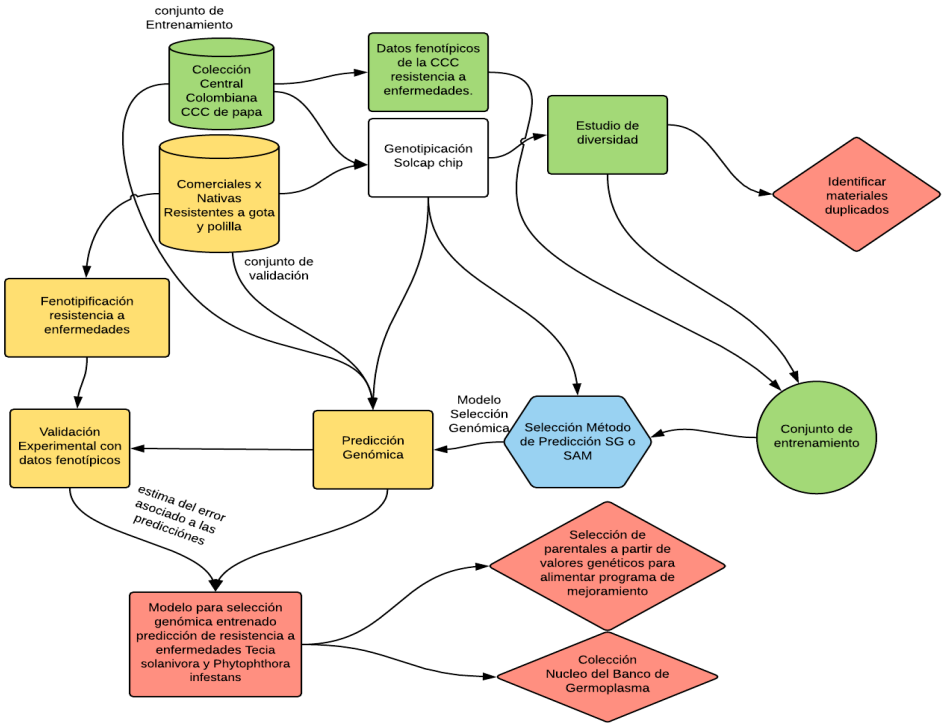

Figura 1. Esquema de la estrategia del análisis de datos del proyecto. En verde se destaca la información utilizada para el conjunto de entrenamiento (**CE**), en amarillo la información genómica usada para hacer la predicción, en azul se destaca el componente en el que se trabajará en esta propuesta y que tendrá un impacto en los bloques rojos que son los productos esperados del proyecto en el cual se enmarca esta propuesta.

Metodología para objetivo 3. Buscaremos características evaluadas para papas diploides y tetraploides en conjunto. Para esto utilizaremos una codificación unificada que permita representar las papas tetraploides (AAAA, AAAB, AABB, ABBB, BBBB) como si fueran diploides (AA, AB, BB) y viceversa. Entrenaremos modelos de predicción genómica utilizando conjuntos con (i) papas tetraploides, (ii) papas diploides y (iii) papas diploides y tetraploides. Posteriormente, evaluaremos las diferencias en el desempeño variando los conjuntos de entrenamiento y mediremos el impacto de utilizar una codificación unificada para ver si es viable utilizar una estrategia con papas diploides y tetraploides simultáneamente.

Cronograma de actividades

OBJETIVO	ACTIVIDAD	Meses											
		1	2	3	4	5	6	7	8	9	10	11	12
1 y 2	Revisión estado del arte. Estudio de los métodos de Predicción Genómica (BLUP, Bayesianos y máquinas de aprendizaje) y de selección asistida por marcadores (GWAS)												
	Conocimiento de la CCC y de los datos fenotípicos disponibles entre estos datos de resistencia a polilla guatemalteca y a gota que se han evaluado en campo y datos de caracterización morfológica												
	Elaboración de flujo de trabajo para cada uno de los métodos de SG y GWAS. Pruebas en el conjunto de entrenamiento.												
	Ajuste de parámetros y medición de desempeño para cada uno de los métodos de predicción. Comparación de desempeño en conjunto de validación.												
	Entrenamiento de modelos SG variando el número de marcadores												
3	Diseñar estrategias para poder usar métodos de SG en papas tetraploides y diploides simultáneamente												
	Elaboración de flujo de trabajo para uso de papas tetraploides y diploides simultanteament e.												

1,2,3	Selección de individuos promisorios para futuros planes de mejoramiento												
	Preparación de manuscrito para publicación en revista A1.												
	Documentación de flujos de trabajo elaborados												
	Socialización de resultados obtenidos con investigadores de otros cultivos para motivar el uso de SG.												

Resultados esperados

	Resultado/Producto	**Indicador**	**Beneficiario**
Generación de conocimiento y/o nuevos productos tecnológicos	Estrategia bioinformática óptima para hacer predicciones de características de interés agronómico en papas nativas a partir de información genómica	Flujo de trabajo con estrategia documentado.	AGROSAVIA y comunidad científica
Fortalecimiento de la capacidad científica nacional	Formación de recurso humano	Investigador con formación Posdoctoral	AGROSAVIA y comunidad científica
Apropiación social del conocimiento	Artículo científico Revista Q1	Manuscrito del articulo	AGROSAVIA y comunidad científica
	Informe final investigación	Archivo impreso y disponible en línea	
	Presentación de resultados a investigadores	Acta de asistencia	

Relación de actividades y resultados a desarrollar por el doctor en el marco de la estancia postdoctoral
El postdoc que trabajará en esta propuesta estará involucrado de manera directa en las diez actividades expuestas en el cronograma de trabajo. Estas actividades culminarán con la elaboración de al menos un manuscrito que resuma todos los resultados y que se someterá a publicación en una revista indexada Q1. Así mismo, esperamos que el postdoc nos ayude a consolidar una línea de investigación fuerte en estrategias de selección genómica en AGROSAVIA que comenzará con este proyecto en papa, pero que esperamos que en el futuro se extienda a otros cultivos.

Financiación – Presupuesto detallado
Este proyecto se desarrollará en la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), Centro de Investigación Tibaitatá (Mosquera-Cundinamarca) en el marco del macroproyecto transversal titulado “Investigación en conservación, caracterización y uso de los recursos genéticos vegetales”. Específicamente, este proyecto se enfoca en apoyar las metas de papa asociadas al proyecto “Diseño e implementación de una plataforma de genotipificación para el BGV” que cuenta con un presupuesto aprobado entre octubre 2018 y octubre 2019 de $355.258.785 pesos colombianos. El investigador postdoctoral tendrá acceso a los equipos, recursos e insumos derivados de proyecto para que pueda desarrollar su propuesta con éxito.

Referencias bibliográficas

1. Departamento Nacional de Planeación. *Documento CONPES 3850. Fondo Colombia en paz. Colombia* (2015).
2. DANE. Boletín técnico-Encuesta Nacional Agropecuaria 2015. 1–25 (2015).
3. Gobierno de Colombia. *Plan Nacional de Seguridad Alimentaria y Nutricional (PNSAN, 2012-2019)*. (2012).
4. Yogendra, K. N. *et al.* Quantitative resistance in potato leaves to late blight associated with induced hydroxycinnamic acid amides. *Funct. Integr. Genomics* **14,** 285–298 (2014).
5. Pelletier, Y., Horgan, F. G. & Pompon, J. in *Insect Pests of Potato* (2013). doi:10.1016/B978-0-12-386895-4.00015-6
6. Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S. & Yockteng, R. Genetic diversity and association mapping in the colombian central collection of solanum tuberosum L. Andigenum group using SNPs markers. *PLoS One* (2017). doi:10.1371/journal.pone.0173039
7. Enciso-Rodriguez, F., Douches, D., Lopez-Cruz, M., Coombs, J. & de los Campos, G. Genomic Selection for Late Blight and Common Scab Resistance in Tetraploid Potato (Solanum tuberosum). *G3 Genes|Genomes|Genetics* **8,** 2471–2481 (2018).
8. Desta, Z. A. & Ortiz, R. Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* **19,** 592–601 (2014).
9. Lorenz, A. J. *et al. Genomic Selection in Plant Breeding. Knowledge and Prospects. Advances in Agronomy* **110,** (2011).
10. Heffner, E. L. *et al.* Genomic Selection for Crop Improvement. *Crops* **49,** 1–12 (2009).
11. Crossa, J. *et al.* Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* **22,** 961–975 (2017).
12. Thorwarth, P., Yousef, E. A. A. & Schmid, K. J. Genomic Prediction and Association Mapping of Curd-Related Traits in Genebank Accessions of Cauliflower. *G3 Genes|Genomes|Genetics* **X,** 1–17 (2017).