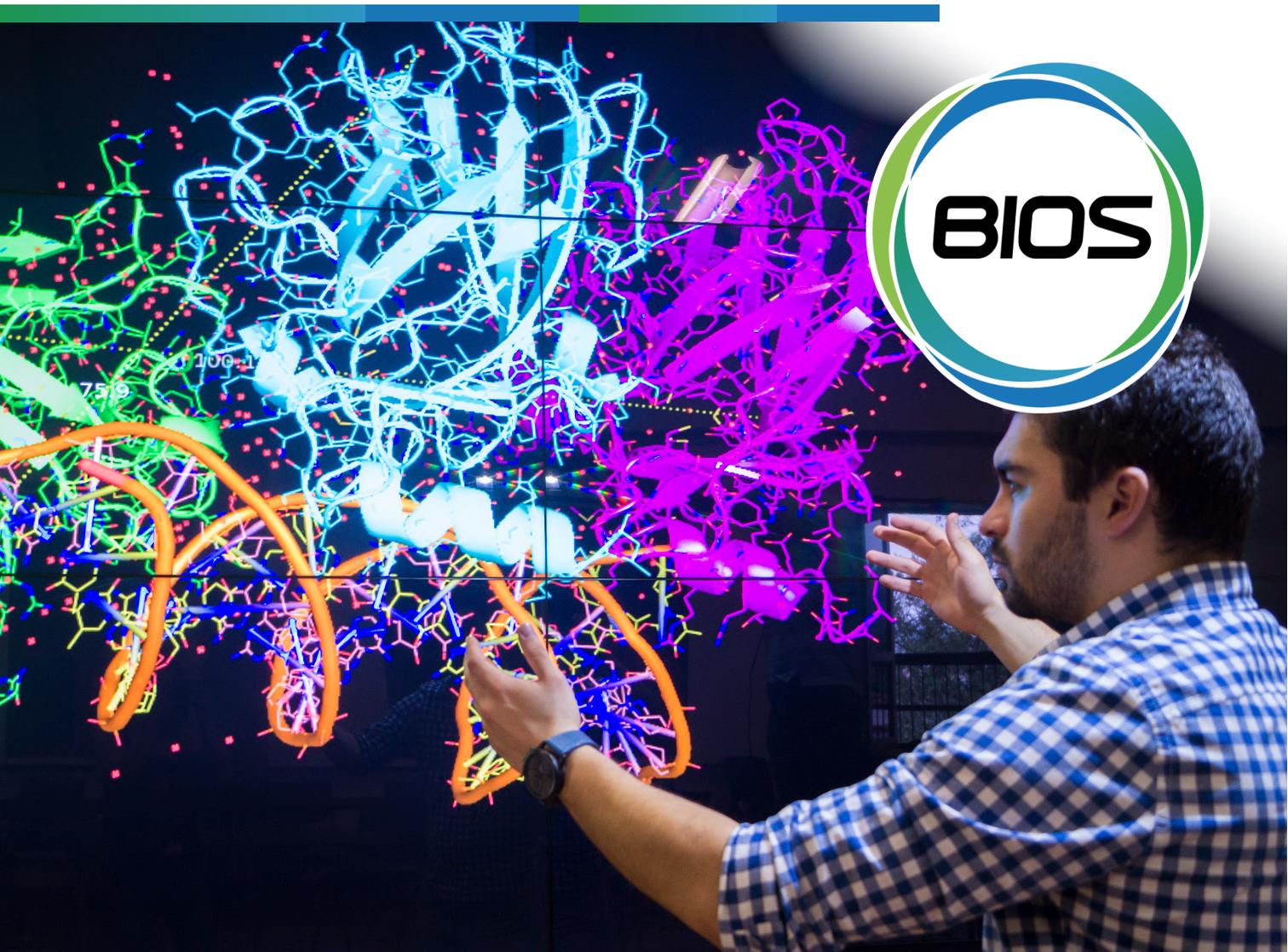


PROYECTOS EMBLEMÁTICOS PARA FORTALECER LA BIOINFORMÁTICA Y LA BIOLOGÍA COMPUTACIONAL EN COLOMBIA

*Ángela María Rojas Sánchez, Marco Aurelio Cristancho,
Dago Hernando Bedoya Ortiz*
Autores



Resultados del encuentro para la identificación de proyectos emblemáticos y conformación de una comunidad Bioinformática



MINTIC



COLCIENCIAS
Ciencia, Tecnología e Innovación

PROYECTOS EMBLEMÁTICOS

Del Centro de Bioinformática y Biología
Computacional de Colombia, BIOS

Editores

Dago Hernando Bedoya Ortiz

Autores

Marco Aurelio Cristancho Ardila

Angela Rojas

Corrección de estilo

Carlos Andrés Urrego

Diseño y gestión editorial

Juliana Grisales Naranjo

Fotografía

Carlos Andrés Urrego

Juliana Grisales Naranjo

ISBN

978-958-59498-0-5



PROYECTOS EMBLEMÁTICOS

Del Centro de Bioinformática y Biología Computacional de Colombia BIOS

Marco Aurelio Cristancho Ardila, Angela Maria Rojas Sanchez, Dago Hernando Bedoya Ortiz
Autores

JUNTA DIRECTIVA BIOS



*David Luna Sánchez
Ministro de Tecnologías de la información y las Comunicaciones*

*Juan David Olarte Torres
Ministerio de Tecnologías de la información y las Comunicaciones*

*Janeth Giha Tovar
Directora de Colciencias*

*Claudia Tinjacá Espinel
Colciencias*

*Marco Casarin Junco
Presidente Microsoft Colombia*

*Carlos Vásquez Gonzales
Microsoft*

*Ricardo Rodríguez Vásquez
Sales Manager Hewlett Packard*

*Jorge Salazar Tovar
Hewlett-Packard*

*Dolly Montoya Castaño
Directora de Investigaciones Universidad Nacional de Colombia*

*Andrés Pinzón Velasco
Universidad Nacional de Colombia*

*Felipe Cesar Londoño López
Rector Universidad de Caldas*

*Luisa Fernanda Giraldo Zuluaga
Universidad de Caldas*

*Luis Fernando Gaviria Trujillo
Rector Universidad Tecnológica de Pereira*

*Diego Mauricio Arias Arango
Director ejecutivo SUEJE*

EQUIPO DIRECTIVO BIOS



Jorge Hernán Gómez Cardona
Director Ejecutivo

Marco Aurelio Crisancho
Director Científico

Dago Hernando Bedoya Ortiz
Coordinador Área de bionegocios y formación



03

INTRODUCCIÓN

10

BIOINFORMÁTICA Y
BIOLOGÍA COMPUTACIONAL

12

HACIA UNA VISIÓN INTEGRADA DE LOS
FENÓMENOS BIOLÓGICOS

15

INVESTIGACIÓN BÁSICA Y APLICADA
PARA LA PRODUCTIVIDAD

20

LINEAMIENTOS DE INVESTIGACIÓN



ÍNDICE

03

BIOINFORMÁTICA EN
BIORECURSOS

10

BIOMEDICINA
COMPUTACIONAL

12

COMPUTACIÓN
CIENTÍFICA

15

DINAMIZADORES

20

PRESTACIÓN DE
SERVICIOS

20

EDUCACIÓN Y
FORMACIÓN

PRÓLOGO

En diciembre de 2014 tuvo lugar en Manizales, el “Encuentro para la Identificación de Proyectos Emblemáticos y Conformación de una Comunidad Bioinformática”, en el marco del convenio suscrito entre COLCIENCIAS y el Centro de Bioinformática y Biología Computacional (BIOS), denominado: “Fortalecimiento del Centro de Bioinformática y Biología Computacional de Colombia” con el objetivo de “Apoyar la investigación científica, desarrollando tecnologías innovadoras que incrementen la productividad del I+D”.

El encuentro se consolidó como un escenario propicio para el planteamiento de ideas, el logro de acuerdos y la toma de decisiones concernientes a iniciativas bioinformáticas con potencial de alto impacto para Colombia; todo en un entorno interdisciplinario del cual hicieron parte diferentes actores y entidades con interés y trayectoria en el tema.

Algunas instituciones como Colciencias, Cenicafé, Cenicaña, CIB, Corpogen, NCBI (National Center for Biotechnology Information), Organizaciones Inteligentes, Roche, Universidad de los Andes, Universidad Nacional de Colombia-Sede Bogotá, Universidad Nacional de Colombia-Sede Manizales, Universidad de Caldas, Universidad Católica de Manizales y Universidad de Antioquia, participaron en esta actividad y expusieron detalles sobre sus respectivas líneas de investigación relacionadas con bioinformática. Asimismo, identificaron las oportunidades de articulación y de generación de convenios con BIOS para potencializar y promover la producción científica.

Los objetivos del evento se enfocaron en: 1) identificar proyectos de alto impacto nacional que representen beneficios multisectoriales, a través de la implementación de herramientas bioinformáticas y biotecnológicas como agentes catalíticos; 2) fomentar la conformación y consolidación de una comunidad académico-científica-empresarial en temas bioinformáticos y biotecnológicos; 3) conocer las iniciativas de formación académica nacional relacionadas con el ámbito de la bioinformática procurando articulación entre ellas y 4) dar a conocer y retroalimentar las líneas de investigación de BIOS.

Uno de los ejes centrales del encuentro fue la participación conjunta de los diferentes actores para formular iniciativas con la potencialidad de incentivar los sectores biotecnológico y bioinformático en el país. Por su viabilidad, impacto y pertinencia, dos de las propuestas presentadas fueron seleccionadas como los proyectos emblemáticos a los cuales debería apuntar la investigación en BIOS. Estas propuestas son El Genoma Colombiano y Expedición Botánica in silico, las cuales se detallan en esta publicación.

PARTICIPANTES

El encuentro para la Identificación de Proyectos Emblemáticos y Conformación de una Comunidad Bioinformática en Colombia, se llevó a cabo el 1 y 2 de diciembre de 2014, en el Recinto del Pensamiento. Manizales, Colombia.

Este encuentro se llevó a cabo en el marco del convenio suscrito entre COLCIENCIAS y BIOS, denominado: "Fortalecimiento del Centro de Bioinformática y Biología Computacional de Colombia" y se enfocó al objetivo de "Apoyar la investigación científica, desarrollando tecnologías innovadoras que incrementen la productividad de I+D".

NOMBRE

Marco Cristancho
Álvaro Gaitán
John Riascos
Simón Buriticá
Marta Lucía Cepeda
Juan Manuel Anzola
Jorge Hernán Gómez
Diego Contreras
Fernando Galeano
Natalia Pabón Mora
Andrés Gómez
Germán López G.
Alejandro Reyes
Germán Castellanos
Emiliano Barreto
Carlos Manuel Estévez
Luis Fernando Castillo
Juan Pablo Montoya
Dago H. Bedoya
Leonardo Mariño
Juan Diego Gómez
Valentina Vargas
Kelly Botero
Andrés Hernández
Miguel Guevara
Patricia Serrano
María Natalia Díaz
Eduardo Gómez
Alejandra Robledo
M. Marcela Velásquez

INSTITUCIÓN

CENICAFÉ
CENICAFÉ
CENICAÑA
COLCIENCIAS
CORPOGEN
CORPOGEN
Organizaciones Inteligentes
ROCHE
Universidad Católica de Manizales
Universidad de Antioquia
Universidad de Antioquia
Universidad de Caldas
Universidad de los Andes
Universidad Nacional de Colombia
Universidad Nacional de Colombia
Universidad Nacional de Colombia
Universidad de Caldas
Clúster del triángulo del café
BIOS
NCBI – BIOS
Facilitadora de espacios creativos
Redacción de relatoría

INTITUCIONES PARTICIPANTES DEL “ENCUENTRO PARA LA IDENTIFICACIÓN DE PROYECTOS EMBLEMÁTICOS Y CONFORMACIÓN DE UNA COMUNIDAD BIOINFORMÁTICA”

CONSIDERACIONES DE LOS PARTICIPANTES SOBRE EL DESARROLLO DE LA BIOINFORMÁTICA EN COLOMBIA

SELECCIÓN DE LOS PROYECTOS EMBLEMÁTICOS



CONTEXTO

INSTITUCIONES PARTICIPANTES DEL ENCUENTRO

La bioinformática y la biología computacional han experimentado un apogeo considerable en los últimos años y han logrado consolidarse, a nivel internacional, como herramientas transversales para la toma de decisiones sobre el manejo de los recursos naturales y para la generación de bio-productos en áreas de investigación y desarrollo a nivel industrial, académico y ambiental. En Colombia, existen diferentes instituciones que desarrollan investigación en dichas áreas y éstas se dieron cita en el "Encuentro para la Identificación de Proyectos Emblemáticos y Conformación de una Comunidad Bioinformática". A continuación se mencionan los principales lineamientos de sus respectivas líneas de investigación.

NOMBRE

Marco Cristancho
Álvaro Gaitán
John Riascos
Simón Buriticá
Marta Lucía Cepeda
Juan Manuel Anzola
Jorge Hernán Gómez
Diego Contreras
Fernando Galeano
Natalia Pabón Mora
Andrés Gómez
Germán López G.
Alejandro Reyes
Germán Castellanos
Emiliano Barreto
Carlos Manuel Estévez
Luis Fernando Castillo
Juan Pablo Montoya
Dago H. Bedoya
Leonardo Mariño
Juan Diego Gómez
Valentina Vargas
Kelly Botero
Andrés Hernández
Miguel Guevara
Patricia Serrano
María Natalia Díaz
Eduardo Gómez
Alejandra Robledo
M. Marcela Velásquez

INSTITUCIÓN

CENICAFÉ
CENICAFÉ
CENICAÑA
COLCIENCIAS
CORPOGEN
CORPOGEN
Organizaciones Inteligentes
ROCHE
Universidad Católica de Manizales
Universidad de Antioquia
Universidad de Antioquia
Universidad de Caldas
Universidad de los Andes
Universidad Nacional de Colombia
Universidad Nacional de Colombia
Universidad Nacional de Colombia
Universidad de Caldas
Clúster del triángulo del café
BIOS
NCBI – BIOS
Facilitadora de espacios creativos
Redacción de relatoría

COLCIENCIAS



COLCIENCIAS
Ciencia, Tecnología e Innovación

Entre las metas generales a las que está apuntando COLCIENCIAS y que podrían ser de interés para una sociedad bioinformática, está la línea de investigación denominada

“Biotecnología y aprovechamiento sostenible de la biodiversidad”. Dentro de ésta pretenden fortalecer la infraestructura tecnológica-científica del país, así como incentivar la formación de recurso humano. Asimismo, cuentan con cinco megaproyectos entre los que se encuentra la creación de una red de bioinformática, genómica y proteómica y esperan disponer de un fondo exclusivo para investigaciones orientadas al área de la biotecnología.

CENICAÑA



Este Centro de Investigación está desarrollando estudios para identificar genes asociados con la tolerancia de la caña de azúcar a la sequía y al anegamiento, ha diseñado un modelo para la identificación de genes candidatos. Por otra parte, también ha puesto en marcha un proyecto GWAS

CENIAFÉ



Ha puesto en marcha el proyecto “Genoma del café”, dentro del cual se han evaluado las secuencias de algunas variedades de la planta, así como de agentes patógenos y un controlador. En cuanto a la broca del café (*Hypothenemus hampei*) y a la roya (*Hemileia vastatrix*), ya se ha completado el 90% de sus

genomas. Del hongo controlador *Beauveria bassiana* se ha completado el 95% y los porcentajes correspondientes a las diferentes variedades de café analizadas son: *Coffea arabica* (80%), *Coffea eugenioides* (80%) y *Coffea canephora* (85%). La entidad empezó a generar datos genómicos y creó su unidad de bioinformática en el año 1999. Actualmente la mayor parte de sus investigaciones están relacionadas con cambio climático y productividad e involucra análisis de recursos genéticos.

CORPOGEN



La Corporación tiene tres enfoques: 1) Investigación, 2) Educación y 3) Productos y servicios.

Cuenta con cuatro grupos de investigación: Ecología microbiana; Genética y microbiología; Bioinformática y Biotecnología. La bioinformática es el eje transversal de todos sus proyectos, los cuales logran aliar investigación con prestación de servicios. Los convenios y asociaciones con otras entidades son estratégicas para complementar la capacidad de cómputo de dicha Corporación.

NCBI (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION)



El Centro Nacional para la Información Biotecnológica integra varias líneas de servicios e investigación, entre las que se encuentran bases de datos públicas y herramientas de biología computacional. En cuanto a proyectos relevantes en el contexto de la bioinformática y la biotecnología se destaca uno sobre identificación precisa de patógenos

utilizando NGS (Next Generation Sequencing). Esta iniciativa cuenta con la colaboración de CDC (Centers for Disease Control and Prevention). El Centro también está participando en el proyecto colaborativo denominado "Chocogen", el cual cuenta con el apoyo de BIOS Y Georgia Tech, y cuyo objetivo general es realizar una caracterización genética de la población del Chocó.

ORGANIZACIONES INTELIGENTES (OI)

Organizaciones Inteligentes

Coaching, Management y Leadership

Esta firma colombiana especializada en Coaching y Liderazgo se enfoca en potencializar el conocimiento y habilidades de equipos de trabajo de manera que puedan generar resultados más eficaces e incrementen la capacidad competitiva de las organizaciones.

Maneja un modelo de sostenibilidad de centros de I+D+i, el cual se basa en varios capitales o activos, entre ellos: Capital humano, capital tecnológico, capital financiero y capital conocimiento.

UNIVERSIDAD CATÓLICA DE MANIZALES



**Universidad
Católica
de Manizales**

El IMBA (Instituto de Microbiología y Biotecnología Agroindustrial) cuenta con algunas líneas de investigación relacionadas con bioinformática, entre ellas seguridad alimentaria y seguridad energética. La primera incluye aprovechamiento de residuos agroindustriales, búsqueda de biocontroladores y promotores de crecimiento vegetal y tratamiento de aguas residuales. La segunda, está orientada a la búsqueda de estrategias para producción alternativa de energía.

UNIVERSIDAD DE CALDAS



La Universidad de Caldas hace parte del proyecto que puso en marcha la Alianza Suma de Manizales para crear la Maestría en Bioinformática y Biología Computacional en la capital de Caldas. Se trata de una iniciativa colaborativa de la que también hacen parte la Universidad Católica, BIOS, la Universidad Autónoma y la Universidad de Manizales. En cuanto a investigaciones relacionadas con bioinformática o biotecnología, actualmente se está llevando a cabo en la Institución un proyecto de genómica de hongos macromicetos con importancia biotecnológica y se prevé iniciar otro relacionado con patógenos del tomate cherry y otro de algas y biocombustibles.

UNIVERSIDAD NACIONAL DE COLOMBIA (SEDE MANIZALES)



**UNIVERSIDAD
NACIONAL
DE COLOMBIA**
SEDE MANIZALES

Uno de los objetivos del "Grupo de Procesamiento y Reconocimiento de Señales" es desarrollar herramientas bioinformáticas orientadas tanto al contexto investigativo como de innovación. Han elaborado aplicaciones para el monitoreo y asistencia en el área médica y otras enfocadas a la interpretación de señales y diagnóstico en casos de epilepsia y enfermedad de Parkinson.

UNIVERSIDAD NACIONAL DE COLOMBIA (SEDE BOGOTÁ)

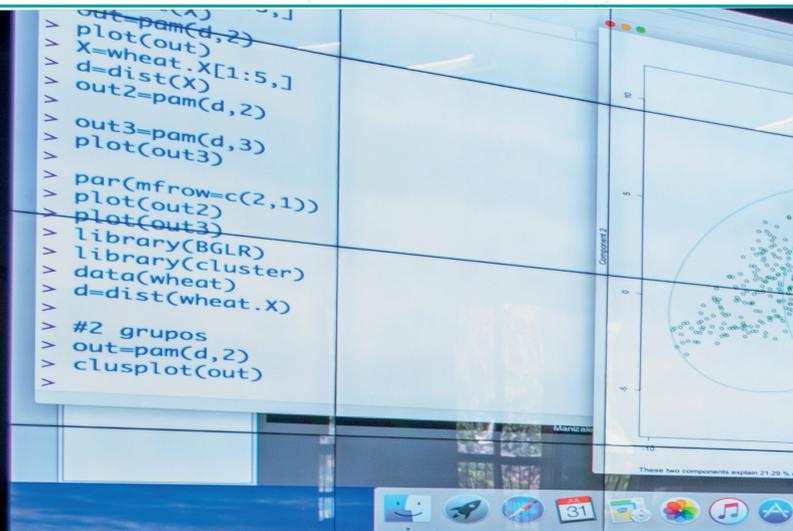


El motor de las investigaciones del Instituto de Biotecnología gira en torno a dos grandes áreas: 1) Bioinformática y salud y 2) Biodiversidad. La primera de ellas incluye el desarrollo de herramientas bioinformáticas orientadas al modelamiento y búsqueda de información para la caracterización molecular de organismos causantes de infecciones nosocomiales. En cuanto al área de biodiversidad, ésta comprende el desarrollo de modelos de análisis para la integración de datos morfológicos, fisiológicos, bioquímicos y moleculares de la biodiversidad colombiana. Actualmente el Instituto también está apuntando a iniciativas relacionadas con la bioprospección.

UNIVERSIDAD DE LOS ANDES



La Institución promueve el desarrollo de herramientas computacionales aplicadas a diversidad microbiana y epigenómica. Asimismo, varias de sus facultades llevan a cabo investigaciones relacionadas con bioinformática orientadas a áreas como genética evolutiva, fitopatógenos, modelaje de proteínas, modelos de membrana, entre otros. También trabajan en la creación de un centro especializado en High Performance Computing (HPC).



ACERCA DE BIOS



Son tres las áreas de investigación del Centro de Bioinformática y Biología Computacional de Colombia: informática en biorrecursos, biomedicina computacional y computación científica. A continuación se presenta información de cada una de ellas, con sus respectivas líneas de investigación y proyectos en marcha.



1. INFORMÁTICA EN BIORECURSOS

Enfocada a la creación y prestación de servicios para el análisis computacional de datos biológicos. “Busca diseñar, implementar, innovar y mantener la plataforma de software para el análisis de datos biológicos, la cual será parte del portafolio de servicios y soluciones de BIOS y que soportará los proyectos de investigación y productividad del Centro”.

Líneas de investigación:

BIODIVERSIDAD Y BIOPROSPECCIÓN

Esta línea tiene como propósitos generar conocimiento aplicable al desarrollo de bio-productos con potencial de comercialización; mejorar la productividad en sectores de importancia económica y propiciar un aprovechamiento sostenible del medio ambiente. Adicionalmente, apunta a la identificación de organismos de interés a partir de proyectos metagenómicos en diferentes ambientes.

BIOCOMBUSTIBLES Y FUENTES ALTERNATIVAS DE RECUROS

Optimización de fuentes biológicas de energía. En ese sentido, incluye indagaciones en torno a aspectos genéticos de especies de algas y microorganismos de importancia biotecnológica, utilizando herramientas bioinformáticas y de biología sintética. Asimismo, involucra investigaciones sobre fuentes de energía renovable y alternativa.



BIODIVERSIDAD Y BIOPROSPECCIÓN

Uno de los motores en torno a los cuales gira esta línea de investigación se centra en los retos que enfrenta el sector agrícola del país ante la amenaza que representa el cambio climático. En ese panorama, BIOS prevé participar considerablemente en iniciativas de investigación sobre adaptación agrícola, no sólo de especies vegetales de importancia económica, sino también de aquellas relevantes para el sector ganadero. Esta línea también acoge el tema de seguridad alimentaria e involucra estudios sobre interacciones planta-patógeno, planta-microorganismos del suelo, animal-patógenos, entre otros.



2. BIOMEDICINA COMPUTACIONAL

Está orientada a la investigación y desarrollo de estrategias conducentes a la prevención, diagnóstico, tratamiento y seguimiento de factores bióticos relacionados con bienestar y salud. Ofrece soluciones computacionales para la vigilancia en salud pública, a través de la permanente recolección, organización, análisis, interpretación, actualización y divulgación de datos relacionados con el sector, para el apoyo a los servicios de diagnóstico médico acertado en el país.

Líneas de investigación:

SERVICIOS E INNOVACIÓN EN SISTEMAS DE VIGILANCIA Y DE APOYO DIAGNÓSTICO

Aborda la generación de mecanismos computacionales de vigilancia en salud pública orientados a desarrollar capacidades de predicción y contención de epidemias, detección temprana de enfermedades, prevención integrada de factores de riesgo y oportunidad de atención de los enfermos. Asimismo, dentro de esta línea se pretende que la infraestructura computacional de BIOS facilite la formación de amplias redes mundiales, nacionales y regionales de expertos para la transferencia rápida y eficaz de conocimientos relacionados con la salud.

CÁNCER Y ENFERMEDADES DE NOTIFICACIÓN OBLIGATORIA

Esta línea apunta a la investigación, a nivel molecular, de enfermedades prioritarias en salud pública como cáncer, tuberculosis, paludismo, dengue, entre otras. También incluye el desarrollo de las herramientas computacionales necesarias para analizar e interpretar la información.



procedente de tales investigaciones. Cabe resaltar que el apoyo de otras instituciones es de especial relevancia para evaluar aspectos como variantes genéticas relacionadas con las patologías de interés, expresión diferencial de genes involucrados con funciones inmuno-moduladoras y de genes virales que puedan alterar las funciones de genes normales. El fin último de esta línea será, entonces, apoyar procesos diagnósticos y hallar mecanismos efectivos de control para diversas enfermedades.

MODELOS DESCRIPTIVOS Y PREDICCIONES EN SALUD PÚBLICA Y EPIDEMIOLOGÍA

El desarrollo de ambos tipos de modelos tiene como objetivo facilitar la evaluación sistemática de las enfermedades y la dinámica de dispersión de ciertos brotes o epidemias. La construcción de los mismos involucra la identificación de mutaciones en los patógenos, de manera que se posibilite conocer la dinámica y cinética molecular de tales agentes causantes de enfermedades para diseñar medidas preventivas y diagnósticas. Finalmente, esta línea incluye la implementación de modelos que permitan evaluar los comportamientos epidémicos durante los últimos veinte años para identificar en ellos posibles factores de riesgo ambientales o genéticos.

MODELOS DESCRIPTIVOS Y PREDICCIONES EN SALUD PÚBLICA Y EPIDEMIOLOGÍA

Profundizar en el conocimiento de aspectos genéticos específicos relacionados con la variabilidad de respuestas a los fármacos, así como a identificar variantes genéticas asociadas con riesgo diferencial. Si bien esta línea aporta herramientas óptimas para la prevención y diagnóstico, su mayor beneficio se concentra en ofrecer a cada paciente el tratamiento más efectivo en función de sus características fenotípicas y genotípicas.



3. COMPUTACIÓN CIENTÍFICA

Es un área transversal que sirve de apoyo a las dos anteriores, convirtiéndose en el pilar investigativo que habilita a BIOS para innovar. Su propósito es dotar a BIOS de capacidades científicas y técnicas en computación de alto desempeño para construir y sostener una posición de liderazgo en bioinformática y biología computacional, desarrollando herramientas especializadas. De esa manera, su accionar se centra en el “uso de Benchmarking y análisis de escalabilidad, desarrollo y validación de arquitecturas y desarrollo de aplicaciones y apoyo a la paralelización”.

Líneas de investigación:

VISUALIZACIÓN DE DATOS

Está relacionada con el desarrollo de técnicas y modelos para la construcción de representaciones visuales de la información. Incluye procesos de adquisición, análisis, integración y transformación de los datos y está orientada tanto a la visualización compleja de estructuras de datos como a herramientas de visualización de Big Data. Actualmente, el Centro cuenta con una infraestructura de visualización robusta e innovadora, que cuenta con 32 pantallas de 55 pulgadas cada una, que entrega una resolución combinada de 66 megapíxeles , equiparable a otros de gran magnitud a nivel mundial.

BIG DATA AND LARGE SCALE MACHINE LEARNING

Se dirige a la generación y adaptación de algoritmos y tecnologías para gestionar datos y modelos predictivos y analizar de forma masiva datos a gran escala (Big Data).



ARQUITECTURAS AVANZADAS DE CÓMPUTO DE ALTO DESEMPEÑO

Involucra el estudio y desarrollo de modelos de cómputo y de hardware y software integrados para facilitar el uso eficiente de los recursos computacionales por parte de la comunidad científica.

CONSIDERACIONES DE LOS PARTICIPANTES

¿CUÁLES SON LAS FORTALEZAS Y DEBILIDADES EN BIOINFORMÁTICA A NIVEL NACIONAL?

Las áreas de Bioinformática y Biología Computacional se encuentran en una fase temprana de desarrollo a nivel nacional, razón por la cual aún no se ha hecho extensivo su uso ni se ha optimizado su desarrollo. Adicionalmente, hay desconocimiento del mercado por parte de los investigadores y de diferentes sectores económicos del país, situación que no ha permitido disponer de una fuente estable de recursos financieros ni de un marco normativo claro concerniente al acceso a los recursos genéticos. Esta situación se ha traducido, por una parte, en la redundancia de esfuerzos de diversas entidades para el desarrollo de líneas estratégicas de investigación y, por otra, en una limitada interacción con el sector productivo. A este panorama se suma la falta de una masa crítica de bioinformáticos, la escasez de plataformas de NGS (Next Generation Sequencing) y la incipiente capacidad de cómputo y conectividad requeridos para propulsar el sector bioinformático.

Sin embargo, pese a tales circunstancias, es importante resaltar que el país sí cuenta con la oportunidad de impulsar un crecimiento significativo de los sectores de Bioinformática y Biología Computacional como punto de partida para el desarrollo de ciencia aplicada. Frente a la creciente demanda de nuevos mercados sostenibles, basados en el adecuado aprovechamiento de los recursos biológicos, ha aumentado paulatinamente el soporte presupuestal para el sector y, aunque poco, se cuenta con personal humano altamente capacitado y con experiencia en el área de Investigación y Desarrollo. Asimismo, la existencia de grupos de investigación con líneas específicas de trabajo, los repositorios de programas bioinformáticas y la oferta académica en el área, representan un motor de crecimiento prometedor.

¿CUÁL ES EL ROLL DE LA COMUNIDAD BIOINFORMÁTICA EN COLOMBIA?

A juicio de los participantes del Encuentro, la Comunidad Bioinformática de Colombia habrá de consolidarse como una herramienta que potencie el desarrollo del país, desde diferentes esferas de acción. Será relevante promocionar su reconocimiento y posicionamiento a nivel nacional e

internacional, así como la asociación con comunidades de otros países, como las de Brasil y Argentina.

Deberá contar con un nodo central que tenga la capacidad de influir sobre políticas públicas para la toma de decisiones en el sector y habrá de propender por el fortalecimiento del Congreso Colombiano de Biotecnología con el fin de convertirse en un punto central de acopio de recursos y servicios. De igual manera, habilitaría la generación de un repositorio de scripts, o una especie de "hackathon" de bioinformática, para captar información de los últimos avances en el tema.

Debe ser integrada y multidisciplinaria para responder a necesidades puntuales sin manifestar redundancia de esfuerzos y propender por la capacitación del recurso humano. Finalmente, deberá contar con el acceso a los recursos genéticos del país, en el marco de una normatividad justa y enfocada al desarrollo científico sin comprometer la conservación de los ecosistemas nacionales.



SELECCIÓN DE LOS PROYECTOS EMBLEMÁTICOS

¿QUÉ PROYECTOS EMBLEMÁTICOS DEBERÍA REALIZAR COLOMBIA EN TEMAS DE BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL?

La restricción del acceso a los recursos genéticos representa un obstáculo real para la puesta en marcha de muchas iniciativas. Sin embargo, no debe descartarse la posibilidad de realizar proyectos a partir de la disponibilidad actual, de manera que pueda responderse a necesidades puntuales de diferentes sectores del país. Con base en esa premisa, los participantes del “Encuentro para la Identificación de Proyectos Emblemáticos y Conformación de una Comunidad Bioinformática”, plantearon, en el marco del evento, una serie de iniciativas concernientes a las áreas de Bioinformática y Biotecnología, las cuales se listan en la Tabla 1.

Una vez expuestas las iniciativas, se procedió a analizarlas teniendo en cuenta su viabilidad, impacto y capacidad para conectar esfuerzos interinstitucionales y se integraron varias de ellas en una única propuesta si presentaban convergencias relevantes.

En el momento de evaluar los proyectos propuestos, cada participante emitió dos votos: uno por viabilidad y otro por capacidad para conectar instituciones. Al final de la actividad, se analizó cuáles proyectos lograron el mayor número de votos y a partir de los resultados se priorizaron los que se presentan en la Tabla 2.

Tabla 1

PROYECTO EMBLEMÁTICO	BREVE DESCRIPCIÓN
Consortio de Genómica	Incluirá la consolidación de dos consorcios: uno de genómica humana y otro de genómica no humana.
El Genoma Colombiano	Su alcance se extenderá a sectores como farmacogenómica, epidemiología, genómica personalizada y diseño inteligente de fármacos.
Expedición Botánica “in silico”	Abarcará genómica de plantas nativas, biología de sistemas, modelos de interacción planta-patógeno y microbiología de suelos.
Ingeniería Metabólica	Comprenderá temas de Biorremediación y Biorrefinerías.
Repositorio Nacional para consulta y uso potencial del recurso genético	Se enfocará en cadenas de producción y correspondería a un “GenBank Colombia”
Caracterización Agroclimática	Incluirá microbiología de suelos y agroecología.
Análisis Orientados a Mejoramiento Genético	Se centrará en la optimización de la Productividad del sector Agrícola.
1000 Transcriptomas o 1000 Genomas	Proyecto a gran escala para analizar los transcriptomas de 1000 especies de plantas.
Microbioma Humano/Nutrigenómica	Referente a la asociación de dietas específicas con el desarrollo de algunos tipos de patologías.

Tabla 1 (continuación)

PROYECTO EMBLEMÁTICO	BREVE DESCRIPCIÓN
Barcoding/Fingerprinting	Propenderá por la creación de certificación de origen a partir de la identificación de la procedencia de determinados organismos, especialmente de aquellos empleados para consumo humano.
Aprovechamiento de medicina tradicional	Se enfocará en el desarrollo de productos con valor agregado a partir del aprovechamiento de la medicina tradicional.
Zoonosis y otras amenazas a la salud pública	Su eje de acción se orientará hacia la prevención y control de determinadas enfermedades.

Tabla 2

PROYECTO	VIABILIDAD	CAPACIDAD de generar conexión interinstitucional
El Genoma Colombiano	11 votos	4 votos
Expedición Botánica in silico	6 votos	12 votos
Desarrollo de productos de medicina tradicional	2 votos	0 votos
Microbioma humano	1 votos	2 votos
Análisis orientado al mejoramiento genético	1 votos	1 votos

El grupo de cinco proyectos se redujo aún más cuando algunos participantes sugirieron que dos de ellos integrarían a los demás. De tal forma, los dos proyectos emblemáticos fueron:

El Genoma Colombiano

Objetivo: Determinar el arsenal genético de las poblaciones colombianas

Temas o subproyectos que incluye: Microbioma humano, farmacología, epidemiología, genómica personalizada, diseño inteligente de fármacos y nutrigenómica.

Expedición Botánica “in silico”

Objetivo: Estudiar 1000 transcriptomas de especies de plantas que puedan tener aplicación en diferentes sectores (industria, cosméticos, agroindustria, energía, salud, entre otros)

Temas o subproyectos que incluye: Expedición botánica in silico, mejoramiento genético, desarrollo de productos de medicina tradicional. Una vez acordados los proyectos emblemáticos de alto impacto a nivel nacional, en términos de bioinformática y biotecnología, se procedió a analizar con qué recursos se cuenta y cuáles harían falta para ponerlos en marcha. La tabla 3 reúne dicha información.

PROYECTO

El Genoma Colombiano

¿CON QUÉ CUENTA?

- Instituciones y grupos de investigación que hacen genómica.
 - Protocolos
 - Datos
 - Procesos
 - Consorcios internacionales
- Algunas iniciativas similares (Proyecto Candela, NatGeo)

¿QUÉ FALTA?

- Apoyo por parte de instituciones médicas
 - Priorizar enfermedades
- Inventario de datos genómicos
 - Accesibilidad a poblaciones alejadas (indígenas)
- Mapear proyectos existentes
- Incluir entidades territoriales
- Consentimientos informados
- Caracterización genómica completa.

Expedición Botánica “in silico”

- Bancos de germoplasma (CENIS, Universidades, CIAT, CORPOICA, Humboldt)
- Bases de datos/ colecciones biológicas (Herbarios, Humboldt)
- Especialistas en taxonomía
- Infraestructura (Equipos para ensamblaje y secuenciación)
 - Aparatos y especialistas de laboratorios (ONG´s, Instituto de Genética UNAL, CIAT, Uniandes, Corpogen)
- Gerencia del proyecto (BIOS podría ser candidato)

- Acceso a recursos genéticos
- Articulación entre entes gubernamentales-academia
- Identificación de capacidades.



PROYECTOS EMBLEMÁTICOS

Uno de los principales retos que enfrentan actualmente las ciencias de la vida es lograr la integración de la cada vez más numerosa información biológica y su interpretación en un contexto sistémico y funcional. La aproximación a un conocimiento biológico integral está circunscrita al desarrollo de métodos basados en el uso de algoritmos y al avance de técnicas y sistemas de computación sofisticada generados a partir de la biología computacional y la bioinformática. Esta última, genera, almacena, organiza, procesa y distribuye un amplio número de datos biológicos por medio de grandes computadores y programas informáticos especializados, de tal manera que su aplicación integral logra reunir información procedente de diferentes áreas del conocimiento, como farmacéutica, medicina, ecología y ciencias agronómicas [1].

Actualmente, los intrincados procesos biológicos que subyacen al funcionamiento de los organismos son abordados, a diferentes escalas, por las ciencias ómicas (genómica, proteómica, metagenómica, transcriptómica, nutrigenómica, epigenómica, entre otras). Estas permiten un abordaje detallado de diferentes mecanismos funcionales dando lugar a la generación de datos y anotaciones biológicas de genes, proteínas, biomoléculas y redes metabólicas. Es notable como el uso de herramientas bioinformáticas cobra cada vez más fuerza en el ámbito de la investigación biológica, facilitando el análisis de la prolija cantidad de datos que se genera actualmente con tecnologías de alto rendimiento, como las de secuenciación de nueva generación (NGS-Next Generation Sequencing). El procesamiento de dicha información a través de programas bioinformáticos permite realizar mapas genéticos aproximados de diferentes organismos de interés a partir de análisis comparativos, como ya se ha hecho en el marco de proyectos de escala global como el del genoma humano y de algunas especies de interés industrial y económico, como el café, el arroz o la palma [1].

Los dos proyectos emblemáticos que se describirán en las próximas líneas involucran, precisamente, la implementación de herramientas bioinformáticas y de biología computacional para potenciar el desarrollo socioeconómico del país y favorecer sectores como el de la salud y la agricultura a partir del uso sostenible de los recursos biológicos.

Bibliografía

[1] Orozco A. 2013. Hacia la sociedad de la información y el conocimiento. Capítulo 6. Nanotecnología y TIC en Costa Rica.



EL GENOMA COLOMBIANO



INTRODUCCIÓN

Las tecnologías y técnicas de análisis de secuenciación genética han experimentado un avance significativo durante los últimos años y cada vez se extiende más su aplicación en diversos ámbitos. En ese sentido, la optimización de la velocidad y la capacidad de almacenamiento han sido cruciales para impulsar las ciencias genómicas. Estas tienen el objetivo de dilucidar aspectos funcionales, estructurales y evolutivos de los genomas con el ánimo de identificar, por ejemplo, patrones de conformación genética que puedan dar información sobre ancestrías o estructura poblacional. La puesta en marcha de este tipo de investigaciones se enfoca, en muchos casos, a la promoción del bienestar humano a través del desarrollo de proyectos de investigación científica con tecnología de vanguardia, formación de recurso humano y generación de aplicaciones genómicas para mejorar la calidad de la atención en salud.

Tras la publicación de los resultados del proyecto “Genoma Humano”, emergieron nuevos retos para la genómica funcional, incluyendo aquellos relacionados con el área de la salud. Entre estos se encuentran, por ejemplo, la detección de factores de riesgo genéticos asociados con la aparición de determinadas enfermedades, el desarrollo de nuevas técnicas de diagnóstico, el diseño de herramientas novedosas para la clasificación de células cancerígenas y, acaso el más importante, la integración de datos genéticos y médicos en la práctica clínica [1].

En ese escenario, se reconoce que los campos de la informática médica y la bioinformática, son idóneos para proporcionar las herramientas necesarias para el cumplimiento de tales objetivos [1]. La bioinformática, por un lado, involucra el desarrollo de novedosas técnicas informáticas aplicables a las ciencias genómicas. Entre tanto, la informática médica es un campo más específico en el marco del cual se diseñan e implementan métodos informáticos útiles para la investigación biomédica. En esta instancia, es importante resaltar que actualmente se cuestiona si existe suficiente conocimiento alrededor de la genómica humana para concentrar esfuerzos significativos en el desarrollo de herramientas que permitan transferir los resultados de la investigación básica a la medicina, y cuáles técnicas médicas e informáticas serían necesarias para realizar dicha transferencia de forma exitosa.

En los párrafos siguientes se hace referencia, precisamente, a la aplicabilidad que se le ha dado en la esfera internacional al conocimiento del genoma humano, poniéndolo al servicio de la salud y el bienestar.

GENÓMICA DE POBLACIONES APLICADA AL ESTUDIO Y TRATAMIENTO DE ENFERMEDADES

Estudio de genes relacionados con cáncer

El cáncer es una enfermedad que ha recibido particular interés en el campo de la investigación biomédica durante varios años. La genómica del éste se encarga de comparar genes provenientes de tumores con genes de tejido sano de una misma persona, para entender el funcionamiento de los que están involucrados en el desarrollo de diferentes tipos de cáncer y tumores, con el objetivo de promover un modelo de medicina preventiva basado en el diagnóstico y detección temprana de la enfermedad para finalmente ofrecerle a los pacientes el tratamiento indicado.

El cáncer inicia cuando las células normales empiezan un proceso de cambio y crecimiento descontrolado, formando una masa conocida como tumor. Estos cambios están dirigidos por los genes, como consecuencia de un error en su información, permitiéndoles a las células cancerígenas la invasión de los tejidos y los órganos sanos. Todos estos genes involucrados en una célula cancerígena son conocidos como el "Genoma del Cáncer", el cual se caracteriza por la presencia de genes con mutaciones específicas responsables de convertir células sanas en células de cáncer. Estas mutaciones pueden ser hereditarias, por azar o en su mayoría se deben a la exposición con factores ambientales y el estilo de vida que lleve la persona. [2]

Uno de los primeros proyectos en aplicar la genómica al estudio del cáncer, fue el análisis del ADN de un paciente con leucemia mielocítica aguda (LMA). Este tipo de cáncer ocasiona que la médula ósea produzca de manera anormal los diferentes tipos de células sanguíneas (glóbulos blancos, glóbulos rojos y plaquetas). Sin embargo, la etiología de la mayoría de los casos de LMA se relaciona con células que se diferencian en glóbulos blancos (exceptuando linfocitos) [3].

El proyecto en cuestión incluyó la secuenciación del ADN proveniente tanto de células sanas como de células anormales y permitió identificar diez mutaciones puntuales en el ADN de las células cancerígenas; dos de ellas ya habían sido previamente asociadas con LMA y las ocho restantes representaron un nuevo hallazgo. En principio, la investigación puso en evidencia la utilidad de técnicas como Whole Genome Sequencing (WGS – Secuenciación Completa del Genoma) y Whole Exome Sequencing (WES – Secuenciación Completa del Exoma) para el descubrimiento de mutaciones novedosas y el posterior desarrollo de terapias potenciales [4]. A partir de este trabajo, surgió el proyecto "Atlas del Genoma del Cáncer" (TCGA), financiado por NIH (National Institutes of Health) en Estados Unidos, cuyo enfoque es el análisis sistemático de marcadores genéticos en más de 20 tipos de cáncer humano a nivel de WGS y WES [4].

Mediante la comparación de muestras de ADN proveniente de individuos sanos y de pacientes, los investigadores buscan identificar marcadores genéticos específicos que puedan asociarse con determinado tipo de cáncer. Asimismo, propenden por realizar estudios genéticos individuales

que permitan arrojar luz sobre respuestas diferenciales al tratamiento [5].

Resulta claro, entonces, que la secuenciación del genoma humano representó un punto de partida clave para ahondar en las bases genéticas que subyacen a la aparición de determinados tipos de cáncer y siguen siendo fundamentales en la búsqueda de alternativas terapéuticas.

Ahora bien, es de destacar que otra de las aplicaciones de la investigación genómica para la detección y tratamiento del cáncer es la posibilidad de hallar indicadores de predisposición familiar, a través de la secuenciación de células germinales. Este tipo de estudios son de gran importancia ya que, en efecto, se estima que la predisposición familiar representa un riesgo del 10% de padecer cáncer de colon, de seno, gástrico y melanoma, y el porcentaje asciende a un 25% para el cáncer de ovarios [4].

Según la Organización Mundial de la Salud, entre 2000 y 2006 se presentaron alrededor de 70.887 casos nuevos de cáncer por año, 32.316 hombres y 38.571 mujeres. Hasta el 2009 se reportaron en Colombia 41.000 muertes asociadas a diferentes tipos de cáncer [5], razón por la cual se ha considerado como una patología de significativa importancia a nivel nacional.

El Plan decenal para el control del Cáncer en Colombia presentado por el gobierno nacional, promueve el bienestar y una vida saludable, basada en la promoción de la salud y prevención de la enfermedad para la detección temprana de los tipos de cáncer de alta prevalencia en el país como medio para fortalecer la capacidad de respuesta del Sistema General de Seguridad en Salud. Se espera entonces reducir la prevalencia de factores de riesgo modificables para cáncer y garantizar la generación, disponibilidad y uso de conocimiento e información para la toma de decisiones en el sector [6]

La detección temprana de enfermedades degenerativas como el cáncer, o incluso el desarrollo de estrategias preventivas a partir de la identificación de factores de riesgo y predisposición, podría beneficiar a miles de personas en el mundo.

El Atlas del genoma del cáncer

Proyecto iniciado en 2005 por el Instituto Nacional del Cáncer y el Instituto Nacional del Genoma Humano en Estados Unidos, con el fin de catalogar cambios moleculares de gran importancia biológica en el desarrollo del cáncer, a partir de la secuenciación genómica y los análisis bioinformáticos. Al comprender mejor la base genética de esta afección a través de la secuenciación, permite a los especialistas diagnosticar, tratar y prevenir la enfermedad a partir de la caracterización, análisis e interpretación de los perfiles moleculares de los tumores de los diferentes tipos de cáncer [7]

Este proyecto inicia como un piloto de tres (3) años en 2006, por parte del Instituto Nacional de Cáncer y el Instituto Nacional de Investigación de Genoma Humano, obteniendo un Atlas sobre diferentes tipos de

cáncer a partir de un convenio entre dos instituciones, desarrollando una infraestructura que permita la fabricación de datos para acceso público, lo cual le facilita a los investigadores validar sus proyectos de investigación en cualquier lugar del mundo [8]

Enfermedades de muy baja prevalencia

Las enfermedades huérfanas se dividen en tres diferentes grupos: huérfana, olvidada y rara. Se dice que una enfermedad es de baja prevalencia cuando afecta a unas pocas personas con relación a la población en general, siendo entonces un problema de salud pública. Por lo general tienden a ser crónicas y de alto riesgo para la vida del que la padece si no es diagnosticada ni tratada a tiempo, esto, a raíz del desconocimiento por parte de los médicos y especialistas. [9]

Según el portal de información de enfermedades raras y medicamentos huérfanos, la Organización Mundial de la Salud (OMS), define una enfermedad de baja prevalencia como aquella que afecta a una persona por cada 10.000 dentro de una población. Se calcula que existen alrededor de 9.300 enfermedades raras y medicamentos huérfanos, los cuales en su mayoría son de origen genético o por epigenética. En Colombia la prevalencia está dada de 20 por 100.000, donde se han identificado alrededor de 13.168 pacientes. [10]

Las técnicas de secuenciación de nueva generación también han resultado útiles para indagar en torno a factores genéticos o epigenéticos asociados con la aparición de enfermedades raras o de muy baja prevalencia. De igual manera, han aportado herramientas para valorar posibles estrategias terapéuticas y para optimizar el diagnóstico clínico de las mismas.

De las más de 7.000 enfermedades raras de carácter monogénico que se calcula afectan a los seres humanos, ya se han identificado los genes para al menos el 50% de ellas. Hoy es posible predecir que con los desarrollos de las nuevas tecnologías de secuenciación, conoceremos el resto de estos genes hacia el año 2020. Para esta misma época, este conocimiento de las enfermedades raras humanas generado gracias a la genética y los mecanismos moleculares que las producen habrán permitido desarrollar terapias para el manejo de la mayoría de ellas.

Aplicaciones de las técnicas de secuenciación para la identificación de factores de riesgo

El diagnóstico de enfermedades puede complementarse con técnicas de biología molecular para la detección temprana de factores de riesgo y predisposición. En ese sentido, ya existen varios proyectos piloto que se han puesto en marcha a nivel internacional y se han creado numerosas bases de datos que acogen una gran cantidad de información puntual sobre genes y secuencias específicas. Algunas de esas iniciativas son:

a. VARIMED (VARiants Informing MEDicine): es una base de datos que contiene los resultados de miles de estudios de epidemiología genética, útiles para deducir las implicaciones clínicas de variaciones específicas del ADN de pacientes [11].

b. OMIM (Online Mendelian Inheritance in Man): se trata de una base de datos que recopila información de genes humanos y desórdenes genéticos. Representa una herramienta importante para la investigación en genética humana y para la práctica de genética clínica [12]. Se puede acceder al recurso electrónico a través de Entrez, el motor de búsqueda de National Library of Medicine.

c. HGMD (Human Gene Mutation Database): es una colección de datos sobre mutaciones de líneas germinales en genes nucleares asociadas con enfermedades humanas hereditarias [13]. Su formato permite la búsqueda de mutaciones simples y aplicaciones avanzadas. HGMD es una herramienta ampliamente usada en el ámbito de la investigación genética.

d. NCBI ClinVar: es un repositorio público de información sobre las relaciones entre variantes genéticas de importancia médica y los fenotipos. Cada registro de la base de datos ofrece información específica sobre el nexo existente entre determinada variante y la salud humana, todo a partir de evidencia científica [14].

CONSIDERACIONES GENERALES SOBRE LOS RETOS DE LA APLICACIÓN DE TÉCNICAS DE SECUENCIACIÓN PERSONALIZADA AL ÁMBITO CLÍNICO

Como se ha mencionado previamente, el avance y desarrollo de nuevas técnicas de secuenciación ha abierto las puertas a un nuevo panorama de investigaciones en torno a factores genéticos asociados con determinadas enfermedades. Sin embargo, es preciso mencionar que a pesar de las bondades que pueda representar el avance de la medicina personalizada, aún existe cierta resistencia por parte de la comunidad médica para incluir el tamizaje genético de cierto tipo de enfermedades dentro del proceso de diagnóstico.

Por otra parte, muchas de las críticas que recibe la aplicación de la secuenciación de genomas o exomas con fines médicos, se enfocan en la insuficiente calidad de los datos para ser transferidos y empleados como herramientas fiables de diagnóstico. En efecto, sí existen aún aspectos por mejorar, como la existencia de falsos positivos. Ante dicha situación, se torna necesaria la creación de plataformas confiables de revalidación, lo que habrá de representar esfuerzos significativos.

Lo cierto es que, al ser el genoma un atributo relativamente estable a lo largo del ciclo de vida de los individuos y al contener el cifrado de la mayor parte de las características del mismo, sí representa un elemento de vital importancia para la detección y diagnóstico de enfermedades. Y en vista de que cada vez avanzan a un ritmo más acelerado las técnicas que permiten adentrarse con mayor profundidad en los detalles del genoma humano, se prevé que la relación costo-efectividad de su secuenciación (y la del exoma e, incluso, el epigenoma), mejore con el tiempo. En ese sentido, se esperan también menores costos y mayor capacidad de procesamiento y almacenamiento de datos. La expectativa es, por consiguiente, que en la medida en que evolucione la tecnología, los datos de la secuenciación alcancen un alto estándar de calidad.

Ahora bien, más allá del avance de las técnicas para la generación de datos genómicos y del desarrollo de aplicaciones para analizarlos, tal vez uno de los principales retos relacionados con la posibilidad de detectar factores genéticos de riesgo, lo representen las enfermedades complejas o multifactoriales. Éstas, a diferencia de las enfermedades mendelianas, son producto tanto de la interacción de genes con factores ambientales, como de eventos epigenéticos y de atributos individuales que no subyacen como tal a la información contenida en el genoma. Es por eso que en patologías de este tipo, como el cáncer, la diabetes, la depresión, el asma, entre otras, la respuesta individual a factores farmacológicos juega un rol determinante [5]. Encontrar marcadores genéticos asociados a estas enfermedades se logra a través de la secuenciación del genoma completo de una muestra significativa de individuos con la enfermedad de interés para compararlo, a nivel genómico, con el de los individuos sin la enfermedad. No obstante, el alto costo de esta estrategia podría representar un verdadero obstáculo, a menos que existiese una fuente de financiación significativa. El proyecto Internacional HapMap se creó, precisamente, para solventar dicho problema.

Proyecto HapMap Internacional

El proyecto internacional HapMap (Mapa Haplotípico), que vio la luz en el año 2002, es una iniciativa internacional que cuenta con el apoyo de varios países, entre ellos Estados Unidos, Canadá, Japón, China, Nigeria y Reino Unido, y cuyo objetivo principal es realizar un mapa haplotípico del genoma humano de manera que puedan determinarse patrones comunes en las variaciones de la secuencia de ADN implicadas en determinadas enfermedades. Provee, por consiguiente, información necesaria para el desarrollo de nuevas herramientas de diagnóstico y para la optimización de las intervenciones terapéuticas [5].

En términos generales, los componentes que se pusieron en consideración para la puesta en marcha del proyecto fueron: selección de las muestras y poblaciones de estudio, consulta pública, selección de SNPs (single-nucleotide polymorphisms – polimorfismos de nucleótido simple), genotipificación, análisis y publicación de resultados [5].

El proyecto se ha desarrollado en tres fases, cada una de las cuales ha permitido aumentar el número de SNPs identificados:

- Fase I (2005): Se genotipificó al menos un SNP común cada 5 kilobases (kb) a lo largo de todo el genoma en 269 muestras de ADN [15].
- Fase II (2007): Se caracterizaron más de 3.1 millones de SNPs, genotipificados en 270 individuos. Adicionalmente, se incluyó entre el 25-35% de las variaciones comunes de SNPs en las poblaciones evaluadas [16].
- Fase III (2009): Se genotipificaron 1.6 millones de SNPs comunes en 1.184 individuos provenientes de 11 poblaciones del mundo. Además, se secuenciaron diez regiones de 100 kb cada una en 692 de los individuos [17].

1000 Genomes Project

"The 1000 Genomes Project", nace en 2008 como una iniciativa para establecer el catálogo más detallado de variaciones genéticas humanas, en una muestra de 1000 participantes anónimos pertenecientes a diferentes grupos étnicos, a través de tecnologías de punta que permitan una mayor agilidad y bajos costos en el proceso de secuenciación. Este proyecto une investigadores interdisciplinarios de China, Italia, Japón, Kenia, Nigeria, Perú, Inglaterra y Estados Unidos, donde cada uno contribuirá a la gran base de datos que mejorará el mapa genético del ser humano. Esta base de datos será de acceso público para la comunidad científica y personas del común. [18]

El objetivo principal es encontrar la mayor cantidad de variantes genéticas que tengan una frecuencia de al menos un 1% dentro de la población estudiada. Esta base de datos será de gran apoyo para el estudio de asociaciones de todo el genoma con estudios en investigación médica.

Diseño del proyecto

Se han realizado 3 proyectos piloto generando información útil para el desarrollo del proyecto de 1000 genomas.

Piloto 1: Baja Cobertura (2008)

Evaluación de la estrategia para compartir datos a través de los datos. Cobertura 2-4X, estrategia Whole Genome Sequencing de 180 muestras. [19]

Piloto 2: Tríos (2008)

Evaluar la cobertura, plataformas y centros. Cobertura entre 20 y 60X. Estrategia Whole Genome Sequencing de 2 tríos (Madre-Padre e hijo adulto). [19]

Piloto 3: Regiones del gen (2009)

Evaluar métodos para captura de la región del gen. Cobertura 50X. 1000 regiones del gen en 900 muestras. [20]

CBBC- BIOS MEDICINA PERSONALIZADA: DIAGNÓSTICO Y TRATAMIENTO PERSONALIZADO

Tradicionalmente, todos los individuos diagnosticados con una enfermedad reciben el mismo tratamiento. Sin embargo, la medicina personalizada consiste en proporcionar un tratamiento único para cada paciente, de manera que pueda serle más afectivo. Para lograrlo, se emplean pruebas diagnósticas específicas y se detectan marcadores biológicos particulares que permitan prever la respuesta del paciente a determinada alternativa terapéutica.

JUSTIFICACIÓN

La investigación genómica en el hombre tiene como objetivo principal el contribuir a la identificación de similitudes y diferencias genéticas en los seres humanos y como esta información tiene un impacto poblacional a nivel socio-económico en términos de salud pública. La investigación y el desarrollo del área biomédica presentan un potencial único en innovación a partir de la prevención, detección temprana y desarrollo de tratamientos especializados de enfermedades asociadas a marcadores genéticos.

La construcción de una base de datos con información genética de la población colombiana para apoyar la provisión de un servicio de salud más incluyente y eficiente en el país, es una de las prioridades de BIOS, consolidándose como una fuente de investigación en casos de éxito donde se involucre la genotipificación en Colombia.

A nivel nacional, BIOS cuenta con la infraestructura de mayor poder computacional (en términos de capacidad de almacenamiento y procesamiento) y con un recurso humano calificado y con experiencia en análisis bioinformático y optimización de algoritmos computacionales para el procesamiento de información biológica. Por lo tanto, cuenta con las condiciones necesarias para desarrollar proyectos de genómica poblacional y detección de enfermedades mendelianas. Por otro lado, al tratarse de una empresa público-privada, cuenta con aliados estratégicos como el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC) y el Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias), patrocinadores del Centro, entre otros, lo cual asegura el requerimiento de una inversión baja para la puesta en marcha de diferentes iniciativas.

Esta propuesta se llevará a cabo junto con la colaboración de aliados estratégicos en el sector salud, con el propósito de crear un catálogo genético de la población colombiana, compuesto por información de diferentes instituciones y especialistas a nivel nacional con un mismo fin, mejorar el sistema de salud nacional y promover la bioinformática y la biología computacional en la investigación de cáncer.

OBJETIVOS



Implementar un “Banco de Perfil Genético” para establecer la susceptibilidad a enfermedades de alta prevalencia en Colombia, a través de cinco enfoques principales:

- 1 Diseñar e implementar una campaña de tamizaje para obtener ADN de una muestra representativa de la diversidad genética de la población colombiana.
- 2 Construir un banco genético con datos que proporcionen información útil sobre la susceptibilidad de la población colombiana a enfermedades de alta prevalencia en el país.
- 3 Desarrollar las herramientas, algoritmos y técnicas necesarias para la genotipificación de la muestra genética obtenida.
- 4 Desarrollar un aplicativo informativo personalizado, basado en métodos preventivos y de estilos de vida saludables.
- 5 Aportar a las decisiones de políticas públicas de salud con base en el análisis poblacional de los resultados.

METODOLOGÍA

Se llevará a cabo bajo tres fases: Genotipificación, Estudios de Asociación a marcadores a una enfermedad de alta prevalencia dentro de la población Colombiana y Divulgación de resultados. En primer lugar será enfocado a estudios asociados a cáncer.

Fase I: Genotipificación

Durante el desarrollo de esta Fase se esperan realizar tres proyectos piloto a nivel regional, con una muestra poblacional de 100 personas por región geográfica. Estos tres escenarios permitirán caracterizar el perfil genómico de la población colombiana a nivel genómico para el desarrollo de estudios para establecer susceptibilidad a enfermedades de alta prevalencia en el país.

Fase II: Asociación de marcadores a enfermedades de alta prevalencia en el país

Durante el desarrollo de esta Fase se esperan realizar tres proyectos piloto a nivel regional, con una muestra poblacional de 100 personas por región geográfica. Estos tres escenarios permitirán caracterizar el perfil genómico de la población colombiana a nivel genómico para el desarrollo de estudios para establecer susceptibilidad a enfermedades de alta prevalencia en el país.

Fase III: Difusión y Expansión

La fase final de difusión estará enfocada en la divulgación de la información para la comunidad científica y del común, con el objetivo de aportar en la toma de decisiones de política en el sector salud colombiano.

Se desarrollará un aplicativo móvil enfocado en la cultura de prevención y mejora del estilo de vida, creando conciencia en las generaciones presentes y futuras del valor de la vida. La información obtenida de los participantes anónimos se manejará bajo términos de privacidad y seguridad de los datos.

La Fase III está compuesta por las siguientes etapas:

1. Socialización de resultados con expertos
2. Campaña de difusión de resultados
3. Desarrollo del Aplicativo Móvil

RESULTADOS ESPERADOS

1. Banco de Perfil genético para futuros estudios médicos y poblacionales.
2. Base de datos de marcadores genéticos relacionados a enfermedades de alta prevalencia en Colombia.
3. Promoción de la medicina preventiva y la detección temprana.
4. Optimizar el sistema de salud del país a través de innovación y tecnología de vanguardia.
5. Aportar en la toma de decisiones de política en el sector salud colombiano.

IMPACTOS

La detección de marcadores genéticos y epigenéticos que puedan asociarse con la etiología de diferentes enfermedades es de vital importancia para optimizar tanto el diagnóstico como el tratamiento de las mismas. En efecto, identificar patrones comunes de variación genética entre pacientes ofrece pistas sobre cómo se relacionan ciertos cambios de secuencia con la velocidad de propagación de la enfermedad y con su repliegue después de un tratamiento [2].

La investigación en genómica representa una oportunidad para encontrar indicios de cómo tratar pacientes con enfermedades de alta o baja prevalencia a futuro. Sin embargo, aunque los datos procedentes de diversos estudios ya están representando cambios en el tratamiento, es muy probable que los pacientes de hoy no se vean beneficiados de los avances en medicina personalizada. Aún queda mucho camino por recorrer para lograr trasladar de manera eficaz los datos genómicos a la esfera clínica. Y precisamente proyectos como el "HapMap Colombia" serán promisorios para materializar el objetivo de ofrecer tratamientos personalizados a los pacientes.

En ese sentido, es de destacar que existe un consenso entre la comunidad científica en que el factor de éxito para que los datos genómicos sean incluidos dentro del diagnóstico y tratamiento de enfermedades, radica en la capacidad de los investigadores de producir aplicaciones exitosas que logren incrementar la precisión y la calidad de dicha información. Por tal razón, un proyecto masivo de tamizaje poblacional contribuirá al avance de la bioinformática y la informática médica para generar datos de altísima calidad.

Por último, debe considerarse que el apoyo del Estado a este tipo de iniciativas será relevante para asegurar que los resultados cuenten con el despliegue suficiente y puedan repercutir favorablemente a nivel nacional.

BIBLIOGRAFÍA

- [1] V. Maojo and C. A. Kulikowski. "Bioinformatics and Medical Informatics: Collaborations on the Road to Genomic Medicine?," Journal of the American Medical Informatics Association, JAMIA, vol.10, no. 6, pp. 515-522, Dic. 2003.
- [2] Organización Mundial de la Salud, Temas de Salud. Cancer 2015 <http://www.who.int/topics/cancer/es/>
- [3] A. C. Society. Leucemia mieloide (mielógena) aguda, 2014.
- [4] E. D. Esplin, L. Oei, and M. P. Snyder, "Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease.," Pharmacogenomics, vol. 15, no. 14, pp. 1771-1790, Nov. 2014.
- [5] The International HapMap Consortium, "The International HapMap Project.," Nature., vol. 18. pp. 789-796, Dic. 2003.
- [6] W. H. Organization. Colombian cancer indicators. Technical report, 2011.
- [7] The Atlas Genome Cancer 2015. <http://cancergenome.nih.gov>
- [8] Plan Decenal para el Control de Cancer en Colombia 2012-2020. Gobierno Nacional.
- [9] Rosselli D., & Rueda J.D. 2011. Enfermedades raras, huérfanas y olvidadas. Colombia.
- [10] Portal de enfermedades raras y medicamentos huérfanos, 2015. <http://www.orpha.net>
- [11] R. Chen, E.V. Davydov, M. Sirota, and A.J. Butte, "Non-Synonymous and Synonymous Coding SNPs Show Similar Likelihood and Effect Size of Human Disease Association.," PLoS ONE., vol. 5, no. 10, pp. 1-6, Oct. 2010.
- [12] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, and V.A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.," Nucleic Acid Res., vol. 33, no. Database issue, pp. D514-D517, Sep. 2005.
- [13] P.D. Stenson, E.V. Ball, M. Morth, A.D. Phillips, J.A. Shiel, N.S.T. Thomas, S. Abeyasinghe, M. Krawczak, and D.N. Cooper, "Human Gene Mutation Database (HGMD): 2003 Update.," Human Mutation., vol. 21. pp. 577-581, Ene. 2003.
- [14] M.J. Landrum, J.M. Leem G.R. Riley, W. Jang, W.S. Rubinstein, D.M. Church, and D.R. Maglott, "ClinVar: public archive of relationships among sequence variation and human phenotype.," Nucleic Acid Res., pp. 1-6, Nov. 2013.
- [15] The International HapMap Consortium, "A haplotype map of the human genome.," Nature., vol. 437. pp. 1299-1320, Oct. 2005.
- [16] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs.," Nature., vol. 449. pp. 851-862, Oct. 2007.
- [17] The International HapMap Consortium, "Integrating common and rare genetic variation in diverse human populations.," Nature., vol. 467. pp. 52-58, Sep. 2010.
- [18] The 1000 Genomes Project Consortium. 2010. An integrated map of genetic variation from 1092 human genomes. Nature 491, 56-65
- [19] The 1000 Genome project Consortium. 2010. A map of human genome from population-scale sequencing. Nature 467, 1061-1073
- [20] Rasmus Nielsen. Genomics: In search of rare human variants. Nature 467 1050-1051







EXPEDICIÓN BIO

INTRODUCCIÓN

Colombia, con más de 54.000 especies registradas en la Infraestructura Mundial de Información sobre Biodiversidad (GBIF), es reconocido por el Programa de las Naciones Unidas para el Medio Ambiente como uno de los 17 países megadiversos que albergan el 70% de la biodiversidad mundial en solo 10% del territorio, y comparte con Brasil el primer lugar a nivel mundial en términos de biodiversidad [1]. Esta extraordinaria característica, además de la relevancia que representa para el equilibrio y bienestar de la biosfera, es considerada ventajosa para potenciar el uso de la biotecnología en beneficio del desarrollo socioeconómico y ambiental del país a partir del manejo y uso sostenible de los recursos naturales, genéticos y sus derivados, mejorando la competitividad nacional a través de la incursión en nuevos mercados de productos innovadores [2].

A escala mundial, la demanda de productos generados a partir del uso sostenible de la biodiversidad, está creciendo mucho más rápido que la demanda de productos tradicionales. Evidencia de ello es que el comercio de estos productos, entre los años 2002 y 2008, en el planeta, experimentó un incremento cercano al 173% [3], lo que impulsó la generación de nuevos mercados especializados de bienes y servicios que promueven el aprovechamiento sostenible y la conservación de los recursos naturales [4]. En consecuencia, a partir de la expansión de los mercados de productos de origen biotecnológico, nació la bioeconomía, un escenario en el que, según la Organización para el Desarrollo y la Cooperación Económica (OECD), toda aplicación tecnológica que utilice sistemas biológicos y organismos vivos o sus derivados para la creación o modificación de productos y/o procesos para usos específicos, contribuye de manera significativa al desarrollo económico de un país, en la medida en que puedan trasladarse a grandes sectores como la industria, la agricultura y la salud [5].

Lo anterior, acompañado de la gran riqueza biológica de los ecosistemas colombianos y del potencial de desarrollo que supone la aplicación de la biotecnología a industrias como bioenergía, biocombustibles, ciencias agropecuarias, biomateriales, ciencias de la salud, cosméticos, biodefensa, biorremediación, conservación, entre muchas otras [6]; plantea el reto de generar frentes de acción para gestionar, de forma sistemática y coordinada, el conocimiento de la biodiversidad y la implementación de nuevas tecnologías para conquistar mercados más sofisticados en los que los productos colombianos puedan marcar la diferencia, al ser innovadores y competitivos [4].

Recientemente, se ha extendido a nivel mundial el término bioprospección. Ésta se relaciona con la búsqueda incesante de recursos, información y

conocimiento por parte del hombre a lo largo de su historia con el fin de promover su bienestar social a través del mejoramiento de la calidad de la salud y la alimentación [7], [9]. Específicamente, el concepto se refiere a la búsqueda sistemática de productos naturales y organismos para su comercialización o uso posterior en procesos de producción en diversos sectores [8]–[10]. De esa manera, se propende por dar un mejor uso y manejo a los recursos naturales, mediante la generación de conocimiento e información útil para optimizar la producción y desarrollo de industrias como la biotecnológica, farmacéutica, cosmética, agropecuaria, química, entre otras.

En efecto, actualmente el uso de recursos biológicos y genéticos como fuente de materia prima para desarrollar productos, constituye una nueva forma de aprovechamiento de la biodiversidad, la cual implica generar el conocimiento y los procesos tecnológicos necesarios para evitar la sobreexplotación de los recursos naturales por acciones antropogénicas [7], [11]. Es por esto que el desarrollo y empleo de tecnologías de punta para la exploración de la biodiversidad con el ánimo de hallar recursos con uso potencial, se convierte en una alternativa para el manejo y uso sostenible de los ecosistemas y para generar bienestar a los colombianos en un contexto de economía verde, el cual busca hacer frente al cambio climático empleando este tipo de tecnologías e industrias en la promoción del crecimiento económico nacional [7], [11].

Ante dicho escenario, las plantas constituyen un invaluable recurso para los procesos bioprospectivos, ya que además del metabolismo primario, presentan un metabolismo secundario a través del cual producen y acumulan compuestos con diferentes propiedades bioquímicas. A estos compuestos, que se denominan metabolitos secundarios o “productos naturales”, se les ha dado un alto valor en los ámbitos médico y económico, debido a los diferentes usos y aplicaciones que pueden atribuírseles en la agricultura, cosmética, farmacia, entre otras [12].

Ahora bien, es de resaltar que los estudios en bioprospección no siempre se realizan para identificar productos comercializables. Los recursos genéticos provenientes de cultivares o hatos y de poblaciones silvestres, representan también uno de los asuntos de mayor interés prospectivo. Las variedades criollas, con sus capacidades de adaptación, rusticidad, resistencia a enfermedades y plagas, son un germoplasma importante en la sostenibilidad de los sistemas agropecuarios mundiales (D. Debouck, comunicación personal, 2002 en [7]). En el estudio realizado por Melgarejo et al. [7], se evidencia que en Colombia los procesos bioprospectivos relacionados con plantas, en relación con las demás taxas, están representados en un 39%, que se distribuye entre grupos de investigación provenientes de universidades, institutos o centros de investigación, empresas, ONGs y Corporaciones Autónomas Regionales. Estas pesquisas abordan la búsqueda de actividades biológicas aprovechables en las plantas a través de la realización de tamizajes o de la evidencia que pueda aportar la etnobotánica, la cual relaciona el conocimiento que tienen del entorno las diferentes comunidades con los usos tradicionales que se les han dado a los recursos vegetales con potenciales económicos.

En el país se cuenta con grandes colecciones que acogen una muestra representativa de la gran diversidad de plantas existentes a nivel nacional. Así, por ejemplo, el Herbario Amazónico Colombiano - COAH del Instituto SINCHI, alberga una colección de aproximadamente 80.000

ejemplares botánicos. De los registros, el 86% están determinados hasta nivel de especie. El Herbario Nacional Colombiano - COL, por su parte, cuenta con 527.000 colecciones de referencia de la diversidad de especies botánicas del país y de sus regiones geográficas.

Frente a los escenarios descritos previamente, este proyecto tiene como propósito explorar las funciones génicas de plantas promisorias de la nación, a través de tecnologías de secuenciación de alto rendimiento y herramientas bioinformáticas, para identificar usos potenciales de la biodiversidad. De esa manera, pretende ser la base sobre la cual se dará continuidad a la Expedición Botánica en la era de la revolución genómica, y para ello se espera contar con el apoyo del Gobierno a través de los ministerios y entidades regionales, ambientales y territoriales, universidades, institutos o centros de investigación, organizaciones empresariales privadas y públicas de interés, Organizaciones No Gubernamentales (ONGs), entre otros.

Antecedentes, alcances y perspectivas de la secuenciación genómica

La secuenciación del genoma de un organismo permite generar el catálogo completo de sus genes, entender los elementos regulatorios que controlan funciones biológicas y posibilita el entendimiento de la variabilidad intra e interespecífica de las especies [13]. Particularmente, la secuenciación de genomas de plantas ha progresado rápidamente desde la publicación, en el año 2000, del genoma de *Arabidopsis thaliana*, una planta herbácea dicotiledónea considerada modelo de otras plantas superiores [14]. Este hecho representó un hito importante tanto en la investigación de plantas como en la secuenciación de genomas, ya que fue uno de los primeros genomas completos de organismos multicelulares en ser secuenciado [15]. Dos años después, fueron publicados los "drafts genomes" (borradores de genomas) de dos subespecies de arroz (*Oriza sativa* L. ssp. *indica* y *Oriza sativa* L. ssp. *japónica*) [16], [17] y en 2005 el "International Rice Genome Sequencing Project" reporta su genoma completo [18]. A partir de allí, se publicaron los genomas de otras plantas como *Brassica oleracea*, en 2005 [19]; *Populus trichocarpa*, en 2006 [20]; *Vitis vinífera*, en 2007 [21]; *Carica papaya*, en 2008 [22], entre otros.

Para la obtención de los genomas de *A. thaliana* y *O. sativa* se utilizaron cromosomas bacterianos artificiales (BACs), donde se clonaron fragmentos de los cromosomas de estas plantas para la obtención de las genotecas de secuenciación [14], [16]–[18]. Por otra parte, los genomas completos de *B. oleracea*, *P. trichocarpa*, *V. vinífera* y *C. papaya* se obtuvieron a través de la tecnología "Whole Genome Shotgun - WGS" de Sanger, conocida como secuenciación de primera generación [19]–[23]. Esta técnica consiste en secuenciar millones de fragmentos del genoma, reduciendo significativamente el tiempo y el costo de la secuenciación de organismos, en relación con la secuenciación mediante el uso de clones en cromosomas bacterianos. No obstante, el uso de esta estrategia de secuenciación puede conllevar un gran margen de error a la hora de ensamblar los genomas, debido a las longitudes cortas de las secuencias obtenidas [15], [24], [25].

Durante varias décadas, la tecnología automatizada de Sanger fue la más utilizada para hacer secuenciación de genomas completos. Sin embargo, algunas desventajas que ésta representa, como el alto coste, el prolongado tiempo de ejecución, el bajo número de reacciones en paralelo que permite realizar y la dificultad del proceso de ensamblaje, hizo necesario el desarrollo de nuevas y mejores plataformas de secuenciación. Fue así como, en la primera década del siglo XXI, surgieron las tecnologías de secuenciación de nueva generación (Next Generation Sequencing - NGS). Estas tienen la capacidad de secuenciar millones de fragmentos de ADN de forma paralela, contribuyendo a la disminución de los costos y aumentando el rendimiento del proceso [23], [26]. La primera de estas tecnologías fue la 454 de Roche, que se basa en pirosecuenciación y que si bien se implementó inicialmente sólo para genomas bacterianos, ha introducido mejoras para extenderse a genomas más complejos [3]. Otra tecnología NGS pionera fue Solexa, ofrecida por Illumina; ésta ofrece mejores rendimientos que 454, pero genera fragmentos mucho más cortos. Otras tecnologías de secuenciación masiva en paralelo son ABI SOLID y Polonator, ofrecidas por Life Technologies y Dover Systems, respectivamente. Todas estas implementaciones tecnológicas fueron denominadas de segunda generación y aunque difieren en varios aspectos, entre ellos la longitud de los fragmentos obtenidos (Figura 1), el esquema principal de trabajo es conceptualmente similar para todas ellas [12], [26], [15].

Figura 1

EQUIPO	COMPAÑIA	SECUENCIACIÓN	DNA MOLDE	LECTURA (bp)	CARRERA (Gb)	TIEMPO CARRERA
GS-FLX (454)	Roche	Polimerasa (pirosecuenciación)	PCR Emulsión	250-400	0.4	10
SOLEXA	Illumina	Polimerasa (terminadores reversibles)	PCR Puente	25-35	1	72
SOLID	Applied Biosystems	Ligasa (octámeros con código de dos bases)	PCR Emulsión	25-35	3	72
HeliScope	Helicos Bioscience	Polimerasa (Extensión asíncrona)	Molécula Única	25-35	14	192
Polonator	Dover Systems	Ligasa (nonámeros)	PCR Emulsión	26	10	80

Figura 1. Tecnologías de Secuenciación.

Tomado de “Aceleración de la Secuenciación Genómica (Rebajas en la secuenciación)”- http://instituto Roche.es/bioinformatica/45/Aceleracion_de_la_Secuenciacion_Genomica_Rebajas_en_la_secuenciacion

Si bien todas estas tecnologías de secuenciación de segunda generación representan importantes avances para la investigación genómica, tienen algunas limitaciones relacionadas, especialmente, con la necesidad de aislar moléculas individuales de ADN en un soporte sólido y con la resolución de las lecturas de nucleótidos obtenidas, las cuales están basadas en microscopía de fluorescencia. Adicionalmente, algunas no pueden generar secuencias superiores a 75 bases, por lo que no son útiles para hacer secuenciaciones de novo. Ante dicha situación,

se ha mantenido un interés constante por desarrollar novedosas técnicas que permitan tanto reducir los costos de secuenciación como aumentar la fiabilidad de los resultados. Precisamente, los avances en nanotecnología y microscopía se han puesto al servicio de la comunidad científica permitiendo el desarrollo y disponibilidad de los denominados secuenciadores de tercera generación, como Ion torrent y PacBio RS systems [24], [28], [29].

Todos estos avances han dado lugar a la revolución genómica, propiciando la convergencia de múltiples áreas del conocimiento relacionadas con biodiversidad, para generar de forma permanente una prolija cantidad de datos que deben ser analizados desde una perspectiva informática, dado el creciente volumen y la complejidad de los mismos [30]. En efecto, el aumento exponencial de datos biológicos ha conducido a la creación de cientos de bases de datos que recopilan y organizan la información, de manera que pueda ser accesible a investigadores de todo el mundo. Por ejemplo, actualmente se han secuenciado y publicado más de 100 genomas de plantas en la base de datos Genbank del National Center for Biotechnology Information – NCBI [31], [32], hecho que ha replanteado el enfoque de cómo se entiende la biología de las plantas y cómo se ejecuta el mejoramiento vegetal [13].

En cuanto a la genómica de plantas, ésta se ha orientado principalmente a aquellas especies cultivadas para la alimentación y nutrición de humanos y otros animales. Este acontecimiento ha impactado positivamente la agronomía, ya que ha permitido la detección de marcadores moleculares asociados con rasgos agronómicos deseables y de genes candidatos dentro de regiones de interés, con los que es posible realizar gestiones eficientes de los recursos genéticos [15], [33], [34].

En Colombia, el Centro Nacional de Investigaciones de Café (CENICAFÉ), con el apoyo del Ministerio de Agricultura y Desarrollo Rural, emprendió en el año 2003 el estudio del genoma del Cafeto (*Coffea arabica* L.), de la broca (*Hypothenemus hampei*) y del hongo entomopatógeno *Beauveria bassiana*. La investigación tenía como propósito enfrentar los retos asociados al cambio climático y aumentar la competitividad de las plantaciones cafeteras mediante el desarrollo de variedades resistentes a plagas y enfermedades. Muchos resultados se derivaron de este macroproyecto y entre los relacionados directamente con el análisis genómico del *C. arabica*, se destacan los mapas genéticos y físicos de su genoma, la identificación del 95% de sus genes, de los cuales algunos se encuentran involucrados en la floración, y la obtención de miles de marcadores moleculares útiles para la evaluación de las poblaciones naturales. Estos resultados permiten determinar asociaciones entre genes (o marcadores) y características agronómicas de interés, como la producción y la calidad, entre otras. Por otra parte, este trabajo ha permitido generar información tanto de los genomas como de las metodologías implementadas, lo que ha favorecido la eficacia de otros proyectos de investigación en torno a esta planta de interés productivo y comercial [35]–[37]

Antecedentes, alcances y perspectivas de los proyectos RNA-Seq

Ante el panorama de la revolución genómica, cabe resaltar que las tecnologías de secuenciación de alto rendimiento han contribuido no solo al crecimiento exponencial de datos genómicos, sino también al desarrollo de nuevas aplicaciones en campos como la transcriptómica y la metagenómica [28]. Particularmente, la transcriptómica es un área esencial para interpretar los elementos funcionales del genoma, en la medida en que ofrece información sobre los componentes moleculares de las redes génicas de diferentes tipos celulares y de la forma en que responden ante determinados eventos [38].

Para poner en contexto, la información del ADN se expresa a través de mecanismos de transcripción y traducción, a partir de los cuales se producen moléculas de ARN y proteínas, respectivamente. La transcripción es un proceso de activación nuclear que depende de estímulos intra o extracelulares para desencadenar cascadas de señalización que determinan cuáles genes deben expresarse o reprimirse de acuerdo al tipo de estímulo inicial. La cantidad de moléculas de ARN mensajero (ARNm) producidas depende de la función que éste tenga en un proceso celular específico. En este sentido, eventos celulares tales como la replicación, la diferenciación y la división celular y otros rasgos fenotípicos, morfológicos, funcionales y de respuesta ante estímulos, son producto de la expresión diferencial de genes a través del proceso de transcripcional [39].

Antes de la aparición y auge de las NGS, se empleaban frecuentemente otras metodologías de genómica funcional, como los microarreglos y la PCR en tiempo real (qPCR, por sus siglas en inglés), para la detección y cuantificación de transcritos, para entender los elementos funcionales del genoma y para medir niveles de expresión génica⁹. La qPCR es una variante de la reacción en cadena de la polimerasa (PCR), que mediante la medición de fluorescencia generada tras la amplificación de productos específicos, permite determinar la cantidad de ADNc (ADN complementario) presente en una o varias muestras. Los microarreglos, por su parte, se basan en la hibridación entre una sonda específica y el ADNc de un organismo en una matriz sólida, habilitando la cuantificación relativa de los transcritos de una célula mediante la intensidad de fluorescencia emitida en la hibridación[40], [62].

Actualmente, la tecnología más empleada para determinar, entre otras cosas, niveles de expresión génica es RNA-Seq. Esta consiste en transformar el RNAm de células expuestas a distintas condiciones ambientales o en diferentes estados de desarrollo, en librerías de ADNc para la posterior secuenciación, a través de tecnologías de alto rendimiento y el respectivo análisis con herramientas bioinformáticas especializadas [40]–[42]. Su implementación ha permitido identificar transcritos y polimorfismos de nucleótido simple (SNPs), obtener perfiles de expresión génica entre diferentes muestras, conocer la estructura de los exones y realizar análisis tanto de expresión diferencial de genes como de patrones de splicing alternativo [38], [40], [42]. La técnica de RNA-Seq ofrece la ventaja de poder cuantificar el nivel de expresión de los genes con mayor precisión y reproducibilidad que a través de microarreglos

y qPCR, en los cuales la medición puede resultar imprecisa debido a que está determinada por una señal fluorescente. Adicionalmente, la RNA-Seq favorece la detección de RNAs no codificantes y de regiones codificadoras de proteínas[40], [63]–[65].

Específicamente, RNA-Seq se basa en las potencialidades de las tecnologías NGS que emplean secuenciación por síntesis (SBS, por sus siglas en inglés), dentro de las cuales la longitud de los fragmentos secuenciados depende del tipo de plataforma implementada, abarcando un rango entre 25 y 450 pb [38], [41], [42]. Los datos de los fragmentos obtenidos son representados por secuencias, reconocidas como lecturas, y la agrupación de éstas genera las llamadas librerías de secuenciación, que además incluyen información particular de las lecturas[43]. Diferentes compañías como Roche/454, Solexa/ Illumina, SOLiD/Life Technologies y Helicos/ BioSciences ofrecen comercialmente la implementación de esta tecnología[44] y mantienen la búsqueda constante de mayores rendimientos de secuenciación y obtención de lecturas más largas a costos cada vez más bajos [45]. Las tecnologías con mayor demanda son Roche/454, porque produce lecturas largas y Solexa/Illumina, porque genera un mayor número de estas a pesar de sus cortas longitudes [39].

Debe mencionarse que independiente de la plataforma a utilizar, un buen diseño experimental para los análisis transcriptómicos es determinante debido a que en los datos de RNA-Seq se encuentran dos niveles de variación; el biológico y el técnico. Por tal razón, si el diseño no se realiza correctamente, se podrían confundir las variaciones técnicas con las biológicas y fácilmente llegar a conclusiones erradas. Además, un error en el diseño experimental sólo podría corregirse reemplazando las muestras con una nueva secuenciación [43], [46]. Para capturar la variabilidad biológica entre tratamientos es necesario realizar réplicas de los mismos; sin embargo, cuando la variabilidad del organismo no se conoce, definir un número adecuado de réplicas biológicas se torna complicado. En estos casos, es pertinente contar con mínimo tres réplicas idénticas por tratamiento para poder evaluar la variabilidad de los organismos dentro de cada uno de ellos , y si es el caso, entre tratamientos [43], [47], [48].

Para analizar genomas completos en una condición dada, es recomendable construir librerías paired-end tags (PET), las cuales secuencian dos regiones por cada fragmento de ADNc, un fragmento a partir de cada extremo, produciendo lecturas pareadas [43], [49]. Esto será importante en el momento de realizar el ensamblaje de los transcritos, ya que al contar con dos lecturas por fragmento aumenta la precisión de la identificación de transcritos hipotéticos y el área de cobertura en el genoma de referencia estudiado, cuando se cuenta con este [43], [50]. Por el contrario, cuando el propósito del estudio es predecir y analizar RNAs no codificantes, se sugiere el uso de librerías single-end, las cuales se obtienen por la secuenciación de un solo extremo del fragmento de ADNc [43], [51].

Otra alternativa para la secuenciación de lecturas es usar la estrategia conocida como barcoding o multiplex, en la cual cada librería contiene patrones de reconocimiento llamados códigos de barra o barcodes (secuencias cortas de alrededor de cinco o seis nucleótidos) contiguos

a la secuencia del adaptador en el fragmento de ADNc, que permiten diferenciar las muestras y hacer más eficiente el uso de cada carril de secuenciación. No obstante, el uso de los barcodes debe ser cuidadoso y se deben tener en cuenta consideraciones importantes en su diseño. Así, por ejemplo, un barcode no puede ser la secuencia reversa complementaria del otro, ni tampoco pueden ser secuencias palíndromas [44].

Ahora bien, una vez obtenidos los datos de RNA-Seq los análisis computacionales que se realizan, cuando existe un genoma o transcriptoma de referencia, consisten en los siguientes pasos fundamentales.

En primer lugar, los datos crudos son convertidos en lecturas de secuencias cortas, las cuales son posteriormente alineadas con la referencia; después, utilizando algoritmos de cómputo, se contabilizan las lecturas mapeadas y se calcula el nivel de expresión de los genes y por último, mediante el uso de herramientas estadísticas, se determina la expresión diferencial de éstos [41]. Ha de resaltarse que no contar con un genoma de referencia no es un requerimiento excluyente de estos análisis, puesto que existen programas capaces de cuantificar la expresión aún cuando no existan anotaciones disponibles; a este proceso se le ha denominado análisis de novo [38], [44].

A las consideraciones anteriores no está de más agregar que en los proyectos RNA-Seq es determinante escoger apropiadamente la plataforma de secuenciación. Cuando el propósito del estudio es realizar ensamblaje de novo, lo más recomendable es obtener lecturas más largas, con el fin de facilitar el ensamblaje. En este caso Roche/454 es una buena opción, aunque es necesario considerar que el número de lecturas obtenidas no será muy alto. Alternativamente, si se cuenta con un genoma de referencia, las lecturas cortas obtenidas con Illumina pueden ser la mejor elección, ya que no es necesario ensamblar el transcriptoma completo y el alto número de lecturas que se generan darán mayor confiabilidad a los datos [52].

Con respecto al proceso de análisis computacional de datos transcriptómicos, es notorio que durante los últimos años se han desarrollado diversas herramientas bioinformáticas como SOAP [53], MAQ [54], RMAP [55], Trinity [56], edgeR [57], Cufflinks [58], entre otras, las cuales brindan a los investigadores la posibilidad de combinarlas para obtener un sistema apropiado de análisis, según los intereses particulares de la investigación [59], [60]. La selección del software a utilizar dependerá, en cierta medida, del número de réplicas biológicas disponibles en el experimento [43].

Como se ha mencionado anteriormente, uno de los objetivos principales al emplear la tecnología RNA-Seq es determinar la expresión diferencial de genes bajo diferentes tratamientos. Para llevar a cabo los respectivos análisis computacionales que permitan lograr dicho objetivo, un paso clave consiste en normalizar los datos de las abundancias de los transcritos hipotéticos. Para ello existen diferentes métodos, dentro de los cuales los más utilizados son la normalización por tamaño de librería y por longitud del fragmento o transcrito [43]. Dos métodos que realizan ambas normalizaciones se conocen como FPKM (Fragments Per Kilobase Of Transcript Per Million Mapped Reads) y RPKM (Reads per Kilobase of Transcript per million mapped reads); la única diferencia entre ambos

es que el primero utiliza fragmentos y el segundo lecturas. Cuando se cuenta con librerías PET suele utilizarse la normalización FPKM, debido a que éstas tienen dos lecturas por fragmento. Una vez normalizados los datos por estos métodos, es posible cuantificar los niveles de transcritos y realizar comparaciones entre muestras [61].

Existen otros parámetros estadísticos que se emplean para analizar las lecturas provenientes de RNA-Seq. Entre ellos se encuentra el fold change de las lecturas, que corresponde a la división del número de lecturas generadas para un gen particular en una muestra vs. el de la otra. La estimación de este parámetro ofrece la posibilidad de correlacionar la expresión de un gen en dos condiciones distintas, así como establecer radios de expresión génica diferencial entre tratamientos [46].

Del mismo modo en que se han introducido cambios para optimizar la secuenciación y análisis de los datos RNA-Seq, este tipo de estudios se ha extendido a diversos tipos de células y tejidos, generando así una cantidad de información sin precedentes sobre las regiones transcritas del genoma, de los extremos 5' y 3' de los genes y de las diversas isoformas de genes conocidos [40]–[42]. Por ejemplo, en el genoma humano, que es uno de los más estudiados, se han confirmado a través de análisis RNA-Seq 31.618 eventos de splicing y se han descubierto otros 379 [66].

MACRO-PROYECTOS INTERNACIONALES A NIVEL GENÓMICO Y TRANSCRIPTÓMICO

1000 PLANTS (1KP)

El proyecto 1,000 Plants (1KP) es un consorcio internacional multidisciplinario que ha generado datos de secuenciación de transcriptomas de aproximadamente 1.000 especies de plantas. Y aunque tiene el potencial de arrojar resultados muy importantes para diferentes ámbitos científicos, su enfoque principal consiste en determinar las relaciones filogenéticas entre especies conocidas e identificar genes asociados con las principales transiciones evolutivas en la historia de las plantas verdes y en el origen de los diversos grupos botánicos. Para la selección de las muestras tuvieron en cuenta proyectos científicos cuyos objetivos pudieran abordarse mediante la secuenciación de múltiples especies de plantas. Adicionalmente, se incluyó el mayor número de linajes del subreino Viridiplantae, el cual comprende a las plantas terrestres (plantas con flores, coníferas, helechos y musgos) y las algas verdes; aunque no todas las especies del subreino se incluyen, por lo menos una especie representa a cada familia [76], [77].

Con el 1KP se han generado aproximadamente 2Gb de datos RNA-Seq por muestra, obtenidos a partir de la plataforma de secuenciación de Illumina (GA2 o HiSeq). En la mayoría de los casos, se secuenció solo un tejido por especie, aunque se incluyeron ciertas excepciones cuando

científico lo requería. Los datos están almacenados en repositorios de WestGrid, los cuales contienen lecturas de transcriptomas ensambladas y sin ensamblar, estimación de niveles de expresión diferencial de genes y promedios de cobertura para Scaffolds. Finalmente, en el marco de este proyecto se secuenció aproximadamente el doble de genes que la totalidad de genes de plantas publicados en las bases de datos públicas. Adicionalmente, a partir del procesamiento y análisis de los datos se han concebido más de treinta publicaciones, entre artículos científicos y de prensa, que han estado liderados por investigadores de diferentes partes del mundo bajo la modalidad de sub-proyectos del 1KP. Se espera, para el 2015, la publicación de otros artículos de los sub-proyectos y de las metodologías empleadas, además de la más importante publicación del 1KP, denominada "capstone" [76], [77].

La financiación de este proyecto fue posible mediante una asociación público-privada; el 75% de los recursos provenían del gobierno de Alberta (Canadá) y el otro 25% del "Musea Ventures". Fue tan grande el impacto del proyecto que científicos de todo el mundo se involucraron en la obtención de las muestras y en la extracción y secuenciación de ARN, además los recursos computacionales fueron proporcionados por el "Compute Canada" y por "China National GeneBank [76], [77].

Plant Genome DataBase Japan (PGDBj)

La PGDBj es un portal web, cuyo objetivo principal es centralizar y organizar los datos y la información genómica de especies de plantas, algas y cultivos importantes para distintos sectores productivos. El portal se enfoca principalmente en los siguientes temas: construcción de una base de datos de genes ortólogos y su integración a las bases de datos de genomas vegetales, integración de las bases de datos de genomas de plantas con marcadores de ADN y su vinculación con mapas de información, integración de las bases de datos de plantas y construcción de un sistema de búsqueda cruzada PGDBj.[78].

De esta forma, la PGDBj ofrece tres bases de datos (DB): 1) DB de ortólogos, la cual proporciona información de clústeres de genes basados en la similitud de las secuencias de aminoácidos; 2) DB de recursos de plantas, la cual integra SABRE DB que provee secuencias de ADNc y genomas almacenados en el RIKEN BioResource Center y el National BioResource Projects; 3) DB de marcadores de ADN, la cual provee información de los marcadores curados manual y automáticamente y la vinculación con mapas de otras bases de datos [78].

JGI Plant Program

El Departamento de Energía del Joint Genome Institute ha llevado a cabo investigaciones en genomas de plantas, algas, hongos y microorganismos, con el propósito de impulsar el desarrollo de biocombustibles. Particularmente, estableció el JGI Plant Program para los proyectos de investigación en plantas, los cuales se han enfocado principalmente en el entendimiento de la fotosíntesis, la conversión de la energía solar a energía química, la caracterización de ecosistemas, el rol de las plantas terrestres y el fitoplancton en la captación de carbono y el papel crucial que juegan las plantas en la descontaminación de suelos afectados con tóxicos o por materias primas para la generación de

biocombustibles. Además, este programa identifica y utiliza metabolitos secundarios para el control positivo y negativo en agricultura [79].

De tal manera, el JGI Plant Program, ha liderado grandes iniciativas de investigación, entre las cuales se destacan:

Phytosome: en este proyecto se realizan estudios de comparación genómica de plantas verdes para comprender relaciones filogenéticas de especies y genes [80].

Plant Flagship Genomes: se planteó con el fin de enfocar los esfuerzos científicos de otros proyectos genómicos y se encarga, específicamente, de seleccionar los genomas de las plantas más importantes para los procesos de desarrollo de biocombustibles [81].

Plant Gene Atlas: el objetivo de esta iniciativa es comprender los efectos de diferentes fuentes de nitrógeno sobre la expresión génica. En ese sentido, pretende construir el catálogo de genes que se expresan bajo diferentes condiciones experimentales en cinco especies (*Panicum virgatum*, *Panicum hallii*, *Sorghum bicolor*, *Brachypodium distachyon* y *Miscanthus sinensis*) escogidas por el Plant Flagship Genomes y tres especies modelo (*Medicago truncatula*, *Phaseolus vulgaris*, and *Zea mays*) [82].

iPlant Collaborative

El iPlant Collaborative es un proyecto que puso en marcha la United States National Science Foundation (NSF) en 2008 para el desarrollo de una ciberinfraestructura que ofreciera capacidades de supercómputo, así como soluciones tecnológicas y sociológicas a investigaciones de gran escala en biología computacional, principalmente a aquellas que tienen el reto de afrontar el crecimiento poblacional, en términos de alimentación, energía y salud. Actualmente, el iPlant Collaborative permite almacenar, compartir y analizar datos biológicos mediante la oferta de poderosos recursos computacionales de software y hardware y brinda apoyo técnico y científico a investigadores estadounidenses y a investigadores de todo el mundo a través de colaboraciones internacionales [83], [84]. Aunque este proyecto no está enmarcado únicamente a investigaciones en plantas, es de particular interés puesto que ha logrado impulsar el trabajo articulado entre científicos y directivos de instituciones como la Universidad de Arizona (UA), el Centro Avanzado de Computación de Texas (TACC), el Laboratorio Cold Spring Harbor (CSHL) y la Universidad de Norte de Carolina en Wilmington (UNCW), para el diseño, desarrollo, implementación y mantenimiento de una ciberinfraestructura que brinde soporte y apoyo constante a investigaciones básicas y aplicadas de las ciencias biológicas enfocadas a los principales retos mundiales [83], [84].

Otras iniciativas internacionales relacionadas con genómica y transcriptómica

Además de los megaproyectos descritos anteriormente, existen otras iniciativas enfocadas a los estudios genómicos de plantas. Para mencionar algunas, se encuentran, por ejemplo: 1001 genomes, para el descubrimiento de variaciones en las secuencias génicas de diferentes

accesiones de *A. thaliana* [85]; el Arabidopsis Información Resource (TAIR), donde se almacenan y mantienen datos de *A. thaliana*, como la secuencia del genoma completo, la anotación estructural y funcional, marcadores genéticos, publicaciones e información [86], [87]; el Sol Genomics Network (SGN), que contiene información genética, fenotípica y taxonómica de los miembros de la familia Solanaceae [88], [89] y la PlantGDB, que contiene datos de secuencias genómicas de plantas dicotiledóneas y monocotiledóneas y desarrolla y provee métodos para la anotación de genomas [90], [91].

Por otra parte, también se han adelantado investigaciones en transcriptomas de cultivos y especies silvestres, los cuales han desempeñado un papel fundamental en la detección de cambios en los niveles de expresión génica de las plantas bajo diferentes condiciones biológicas o en diferentes tipos de células y tejidos, influenciando fuertemente el entendimiento de cómo los factores ambientales ejercen impacto sobre los estadios de desarrollo y las funciones celulares [41], [67].

Otros estudios han permitido identificar y cuantificar transcritos nuevos y conocidos, relacionados con procesos bioquímicos como la fotosíntesis, la asimilación de nitrógeno y la síntesis de metabolitos de interés productivo y/o comercial [68]–[70]. Asimismo, se han identificado genes relacionados con el estrés físico (hídrico y calor) y/o biológico, ofreciendo información útil para comprender los mecanismos que se activan durante diversos estímulos y respuestas inmunes [39], [71].

En cuanto a RNA-Seq, son muchos los proyectos que se han llevado a cabo en plantas. A continuación se describen algunos de ellos con el fin de ejemplificar las múltiples aplicaciones que esta tecnología brinda a la ciencia contemporánea.

- En relación con el entendimiento de procesos bioquímicos usando secuenciación NGS, se realizó un estudio transcriptómico en los diferentes estadios de desarrollo y tejidos de *Setaria viridis*, la cual es considerada planta modelo de la fotosíntesis C4. Los resultados demostraron que la mayoría de genes secuenciados son responsables de regulaciones biológicas y metabólicas, reportándose un total de 25.340 genes distribuidos así: procesos biológicos, 46.28%; funciones moleculares, 15.93% y componentes celulares, 37.79%. Adicionalmente, usando RT-qPCR confirmaron 16 genes relacionados con rutas metabólicas de biosíntesis de terpenoides y mostraron expresión diferencial entre diferentes rutas. Por otro lado, los datos de secuenciación demostraron ser un insumo efectivo para el descubrimiento de marcadores moleculares y la identificación de nuevos genes [68].
- Otro estudio transcriptómico dentro del cual se emplearon tecnologías de secuenciación de alto rendimiento, fue llevado a cabo en cuatro estadios de desarrollo de las flores de *Phyllostachys edulis* (bambú moso). En este trabajo mapearon 18.309 genes en la base de datos Gene Ontology (GO) y 23.522 genes en 302 rutas metabólicas de KEGG. Los genes anotados en GO mostraron niveles de expresión significativamente diferentes, por lo que se sugirió que podrían participar en procesos y rutas relacionadas con el desarrollo de la flor del bambú. Particularmente se identificaron 80 genes relacionados con el desarrollo

floral, los cuales mostraban variaciones de expresión a través del tiempo. Por ejemplo, los genes Dof disminuyeron considerablemente su expresión en la antesis y el desarrollo embrionario. Por el contrario, genes como LHY y PIF3 se sobre-expresaron a medida que avanzaban los estadios florales, sugiriendo que los Dof juegan un rol fundamental en las primeras etapas del desarrollo floral y los LHY y PIF3 en la etapas finales. En conclusión, los resultados revelaron genes clave en la expresión de ciertas propiedades del desarrollo floral en bambú [70].

- En *Litsea cubeba*, una planta aromática de la cual se extraen aceites esenciales con aplicaciones médicas y económicas, se habían reportado hasta el 2013 alrededor de 40 secuencias de nucleótidos en las bases de datos públicas. Sin embargo, tales secuencias no eran suficientes para entender los mecanismos moleculares de biosíntesis. Por tal motivo, mediante tecnologías de secuenciación de alto rendimiento, se emprendió la investigación de su transcriptoma. El estudio permitió identificar una gran cantidad de transcritos y descubrir genes relacionados con estos mecanismos, contribuyendo así al entendimiento de los procesos bioquímicos concernientes a la biosíntesis de los principales constituyentes de los aceites esenciales [54].

- En el campo de estudio de interacción planta-patógeno y de mecanismos de respuesta inmunológica, se lograron obtener en el 2011 análisis globales de transcriptomas de *A. thaliana* [72], *Glycine max* (soya) [73] y *Gossypium barbadense* (algodón) [74]. En el trabajo enfocado en *A. thaliana*, fueron secuenciados y ensamblados múltiples genomas fenotípicamente diversos de diferentes regiones geográficas y dos réplicas biológicas del transcriptoma de la plántula. En este estudio anotaron cada genoma independientemente de la referencia y generaron un catálogo extenso de polimorfismos. Además, el análisis profundo de los datos reveló que la variación en los niveles de expresión génica es mucho mayor en genes que están involucrados en respuestas a estrés biótico. Del mismo modo, identificaron que los genes de resistencia y los relacionados con defensa codifican para proteínas más variables que aquellas codificadas por genes del metabolismo basal [72].

- En algodón, con RNA-Seq y qPCR, se realizó un estudio de su transcriptoma después de ser inoculado con el hongo patógeno *Verticillium dahliae*. En este proyecto se analizaron los perfiles de expresión en raíces a las 4, 12, 24 y 48 h post-inoculación (hpi). Se detectó expresión diferencial entre plantas resistentes y susceptibles en más de 3.000 genes, lo que permitió profundizar en el entendimiento de cómo algunos genes están involucrados en eventos de respuesta inmune [74].

- En soya se realizó un estudio de mecanismos de defensa. Particularmente, se analizó la expresión diferencial de líneas isogénicas susceptibles y resistentes a la bacteria *Xanthomonas axonopodis* pv. *glycines* (Xag). Un total de 46.367 genes fueron mapeados con la referencia, de los cuales 134 exhibieron diferencias significativas de expresión entre líneas susceptibles y resistentes. Los resultados evidenciaron que dichos genes juegan un rol fundamental en la modulación de los mecanismos de respuestas a estímulos abióticos ejercidos por la bacteria [73].

Ahora, en el ámbito nacional, CENICAFÉ ha sido líder en el desarrollo de investigaciones genómicas. Este centro de investigación ha realizado estudios moleculares del transcriptoma de tres tejidos diferentes (hoja, flor y fruto) de *C. arabica* var. Caturra, con el fin de identificar los genes de importancia agronómica que se expresan en estos tejidos. Los resultados han arrojado un total de 1.824 transcritos para tres genotecas, de las cuales 371 fueron anotados en GO, principalmente en las categorías de organización celular y biogénesis, metabolismo, síntesis de proteínas, transporte, transcripción, transducción de señales y respuestas a patógenos o plagas. Esta distribución de categorías se observa también en otras plantas, por lo que en este trabajo concluyen que es una tendencia general para todas las plantas. Por otra parte, se identificaron transcritos relacionados con procesos de respuesta a patógenos y plagas, a estrés abiótico y a resistencia debida a toxinas; hallazgos que pueden ser valiosos para los programas de mejoramiento genético del café [75].



Propuesta del Centro de Bioinformática y Biología
Computacional CBBC-BIOS

CBBC- BIOS 1.000 TRANSCRIPTOMAS – EXPEDICIÓN BOTÁNICA “IN SILICO”

Esta iniciativa surge en respuesta a la necesidad de apoyar el proyecto “Expedición BIO”, del Gobierno Nacional, para recopilar información genética relevante sobre las especies de plantas nacionales de importancia agrícola. Se pretende identificar genes de interés para optimizar la producción agrícola y, por otra parte, promover la bioprospección y conservación de la biodiversidad. La iniciativa será el punto de partida para la creación de nuevos mercados y el mejoramiento de los mercados existentes dentro de un contexto sostenible y rentable a partir de soluciones de innovación científica.

JUSTIFICACIÓN

La revolución tecnológica que ha permeado durante los últimos años a las ciencias de la vida, se conoce como la era del “big data”. Ésta ha permitido la generación de una gran cantidad de datos que deben ser analizados a través de poderosas herramientas computacionales y que aportan luz al entendimiento de diversos procesos biológicos, a diferentes escalas. Dicho conocimiento, a su vez, puede ser aplicado para responder a necesidades específicas de la población humana a nivel global y para generar estrategias de uso sostenible de los recursos naturales. De esa manera, por ejemplo, la investigación genómica de plantas es fundamental para hacer frente a los desafíos que suponen para la agricultura tanto el crecimiento poblacional como los efectos del cambio climático y ofrece información de base para la búsqueda de fuentes de energía renovable. Por consiguiente, plantear iniciativas enfocadas a la generación de conocimiento sobre aspectos genéticos de diferentes especies de interés, habrá de ser prioritario en el contexto nacional.

BIOS cuenta con la infraestructura y el recurso humano necesarios para poner en marcha este proyecto de gran magnitud a través del cual se espera desarrollar una colección biológica de genes de interés que pueda ser de utilidad para fortalecer los sectores agrícola, farmacológico, energético e industrial y para impulsar una economía verde en Colombia a través del uso de la bioinformática y la biología computacional.

OBJETIVOS



Generar conocimiento básico sobre la estructura y función del genoma de plantas de interés, con el fin de fortalecer el desarrollo del sector agrícola, la bioprospección y la conservación de la biodiversidad colombiana a través del acceso a los recursos genómicos y el desarrollo de ciberinfraestructura para estudios de genómica en plantas. Los enfoques específicos del proyecto serán:

1

La re-secuenciación de genomas y transcriptomas de plantas económicamente importantes en Colombia.

2

La secuenciación de novo de genomas y transcriptomas de plantas promisorias, para la identificación de genes de importancia para los sectores farmacéutico, energético, industrial y alimenticio. En esta instancia será vital el apoyo por parte de centros de investigación, universidades y Corporaciones Autónomas Regionales.

3

La obtención de recursos genéticos de plantas con prioridad para la conservación o representativas de la diversidad colombiana, entre ellas plantas endémicas, especies sombrilla y especies amenazadas o en vía de extinción.

4

La secuenciación de novo de genomas y transcriptomas de plantas colombianas que aporten al avance de la ciencia en un contexto global.

METODOLOGÍA

Se llevará a cabo bajo tres fases: Genotipificación, Estudios de Asociación a marcadores a una enfermedad de alta prevalencia dentro de la población Colombiana y Divulgación de resultados. En primer lugar será enfocado a estudios asociados a cáncer.

Fase I: diseño

Se contará con un grupo interdisciplinario el cual estará a cargo un líder especializado en botánica y transcriptómica. El equipo estará conformado por bioinformáticos, ecólogos, ingenieros ambientales, biólogos y microbiólogos, quienes procesarán y analizarán los datos.

Por otra parte, se buscarán alianzas estratégicas para crear una red de información que propicie el uso y aplicación de los productos generados por el proyecto. Los aliados potenciales serían el Gobierno Nacional, a través de los Ministerios de Ambiente y Desarrollo Sostenible, Ministerio de Agricultura y Ministerio de Tecnologías de la Información y las Comunicaciones; entidades territoriales, instituciones educativas y universidades de Caldas, los CENI, las pequeñas agremiaciones y federaciones agrícolas, el Instituto Humboldt y la comunidad involucrada en el conocimiento tradicional asociado a la conservación y uso sostenible de la biodiversidad.

El Diseño Experimental se basará en la recolección de material vegetal para la obtención del recurso genético y derivados de la biodiversidad, todo bajo los lineamientos legales vigentes con fines de investigación científica, no comercial, para un futuro enfoque en bioprospección. El recurso genético será utilizado para desarrollar estudios de transcriptómica comparativa y para generar una plataforma adecuada para los usuarios en la cual se alojará información importante para promover crecimiento industrial, científico y económico del país.

El proyecto se enfocará en aquellas especies de plantas que puedan enmarcarse en uno de los siguientes grupos, cada uno de los cuales implicará alianzas con diferentes entidades en función de los objetivos.

- Plantas cultivables económicamente importantes en Colombia: alianzas con los Centros Nacionales de Investigación y universidades.
- Plantas promisorias o con potencial de impulsar la economía (las plantas medicinales y aromáticas): alianzas con gremios agrícolas e investigadores del sector farmacéutico.
- Plantas con importancia para la conservación de la biodiversidad nacional, en el escenario de cambio climático y adaptación de las especies vegetales del país: alianzas con los institutos de investigación en biodiversidad nacional, como el Humboldt y el Sinchi.

A continuación se presenta la ruta metodológica que se seguirá para los diferentes tipos de plantas sobre los cuales se enfocará el proyecto.

- Recolección de recurso genético con base en las necesidades de los gremios agrícolas para apoyarlos en el área de la bioinformática y la biología computacional.
- Asociación con los Centros Nacionales de Investigación y/o universidades para llevar a cabo, de forma exitosa, todos los procesos moleculares, desde la extracción de RNA hasta la preparación de las muestras para el envío a secuenciación.

Fase II: Desarrollo de plataformas de información

Haciendo uso de la infraestructura con que cuenta BIOS, se desarrollará una plataforma para la visualización de los datos obtenidos, así como diferentes aplicativos enfocados a las necesidades específicas de los usuarios.

Fase III: Sociabilización de resultados y promoción

Se generarán los escenarios necesarios para socializar los resultados del proyecto con los aliados estratégicos y otros actores interesados. Asimismo, se propenderá por buscar soluciones a las necesidades de los usuarios finales, los cuales incluyen investigadores, técnicos de extensión rural y los propios agricultores, dentro de un contexto de innovación científica aplicada.

RESULTADOS ESPERADOS

- Base de datos de la biodiversidad colombiana para generar análisis de transcriptómica comparativa
- Soluciones enfocadas a la optimización del sector agrícola
- Desarrollos biotecnológicos para la creación de mercados sostenibles y para potencializar el sector farmacéutico y cosmético a través de la bioprospección

BIBLIOGRAFÍA

- [1] I. A. Von HUMBOLDT, "La biodiversidad y los servicios ecosistémicos." [Online]. Available: <http://humboldt.org.co/biodiversidad>. [Accessed: 05-Jun-2014].
- [2] "Consejo Nacional de Política Económica y Social - CONPES 3697 Política para el Desarrollo Comercial de la Biotecnología a partir del uso Sostenible de la Biodiversidad," 2011.
- [3] D. N. de P.- DNP, "Colombia adopta política para el desarrollo comercial de la biotecnología.," 2011.
- [4] A. P. Arango, G. De Análisis, and J. D. Vargas, "Plan Estratégico Nacional de Mercados Verdes Ministerio del Medio Ambiente," no. 8, 2002.
- [5] O. for E. C. and Development, "The Bioeconomy to 2030 Main Findings and Policy Conclusions," 2009.
- [6] L. Walker, "Building on Global Drivers of Innovation," *Ind. Biotechnol.*, vol. 8, no. 4, pp. 163–163, 2012.
- [7] C. B. y C. R. MELGAREJO, L. M., J. SÁNCHEZ, A. CHAPARRO, F. NEWMARK, M. SANTOS-ACEVEDO, "Aproximación al estado actual de la bioprospección en Colombia." *Cargrapics*, p. 334, 2002.
- [8] S. A. LAIRD, *Introduction: equitable partnerships in practice.*, Sarah A. L. London: Biodiversity and Traditional Knowledge. *Equitable Partnerships*, 2002, pp. xxii–xxxvi.
- [9] O. D. Torres and L. Velho, "La bioprospección como un mecanismo de cooperación internacional para fortalecimiento de capacidades en ciencia y tecnología en Colombia," *Ciência da Informação*, vol. 38, no. 3, pp. 96–110, Dec. 2009.
- [10] J. M. Feinsilver, "Prospección de la biodiversidad: potencialidades para los países en desarrollo," *CEPAL*, vol. 60, pp. 111–128, 1996.
- [11] L. U. Z. M. Melgarejo, "Bioprospection : National Plan and Up to Date Situation in Colombia," *Acta biológica Colomb.*, vol. 8, no. 2, 2003.
- [12] A. Ávalos and G. Elena, "Metabolismo secundario de plantas," *Reduca Biol. Ser. Fisiol. Veg.*, vol. 2, no. 3, pp. 119–145, 2009.
- [13] C. Feuillet, J. E. Leach, J. Rogers, P. S. Schnable, and K. Eversole, "Crop genome sequencing: lessons and rationales.," *Trends Plant Sci.*, vol. 16, no. 2, pp. 77–88, Feb. 2011.
- [14] The Arabidopsis Genome Initiative, "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.,," *Nature*, vol. 408, no. 6814, pp. 796–815, Dec. 2000.
- [15] M. E. Bolger, B. Weisshaar, U. Scholz, N. Stein, B. Usadel, and K. F. X. Mayer, "Plant genome sequencing - applications for crop

improvement.," *Curr. Opin. Biotechnol.*, vol. 26, pp. 31–7, Apr. 2014.

[16] J. Yu, S. Hu, J. Wang, G. K.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, and H. Yang, "A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*).," *Science*, vol. 296, no. 5565, pp. 79–92, Apr. 2002.

[17] S. A. Goff, D. Ricke, T.-H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B. M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W. Sun, L. Chen, B. Cooper, S. Park, T. C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, and S. Briggs, "A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*).," *Science*, vol. 296, no. 5565, pp. 92–100, Apr. 2002.

[18] International Rice Genome Sequencing Project, "The map-based sequence of the rice genome.," *Nature*, vol. 436, no. 7052, pp. 793–800, Aug. 2005.

[19] M. Ayele, B. J. Haas, N. Kumar, H. Wu, Y. Xiao, S. Van Aken, T. R. Utterback, J. R. Wortman, O. R. White, and C. D. Town, "Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis*," pp. 487–495, 2005.

[20] G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G.-L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Déjardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehling, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjärvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J.-C. Leplé, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouzé, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C.-J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer, and D. Rokhsar, "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).," *Science*, vol. 313, no. 5793, pp. 1596–604, Sep. 2006.

- [21] O. Jaillon, J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pè, G. Valle, M. Morgante, M. Caboche, A.-F. Adam-Blondon, J. Weissenbach, F. Quétier, and P. Wincker, "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.," *Nature*, vol. 449, no. 7161, pp. 463–7, Sep. 2007.
- [22] R. Ming, S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J. H. Saw, P. Senin, W. Wang, B. V Ly, K. L. T. Lewis, S. L. Salzberg, L. Feng, M. R. Jones, R. L. Skelton, J. E. Murray, C. Chen, W. Qian, J. Shen, P. Du, M. Eustice, E. Tong, H. Tang, E. Lyons, R. E. Paull, T. P. Michael, K. Wall, D. W. Rice, H. Albert, M.-L. Wang, Y. J. Zhu, M. Schatz, N. Nagarajan, R. a Acob, P. Guan, A. Blas, C. M. Wai, C. M. Ackerman, Y. Ren, C. Liu, J. Wang, J. Wang, J.-K. Na, E. V Shakirov, B. Haas, J. Thimmapuram, D. Nelson, X. Wang, J. E. Bowers, A. R. Gschwend, A. L. Delcher, R. Singh, J. Y. Suzuki, S. Tripathi, K. Neupane, H. Wei, B. Irikura, M. Paidi, N. Jiang, W. Zhang, G. Presting, A. Windsor, R. Navajas-Pérez, M. J. Torres, F. A. Feltus, B. Porter, Y. Li, a M. Burroughs, M.-C. Luo, L. Liu, D. a Christopher, S. M. Mount, P. H. Moore, T. Sugimura, J. Jiang, M. a Schuler, V. Friedman, T. Mitchell-Olds, D. E. Shippen, C. W. dePamphilis, J. D. Palmer, M. Freeling, A. H. Paterson, D. Gonsalves, L. Wang, and M. Alam, "The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus).," *Nature*, vol. 452, no. 7190, pp. 991–6, Apr. 2008.
- [23] M. L. Metzker, "Sequencing technologies - the next generation.," *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, 2010.
- [24] S. Rounsley, P. R. Marri, Y. Yu, R. He, N. Sisneros, J. L. Goicoechea, S. J. Lee, A. Angelova, D. Kudrna, M. Luo, J. Affourtit, B. Desany, J. Knight, F. Niazi, M. Egholm, and R. a. Wing, "De Novo Next Generation Sequencing of Plant Genomes," *Rice*, vol. 2, no. 1, pp. 35–43, Mar. 2009.
- [25] R. Li, J. Ye, S. Li, J. Wang, Y. Han, C. Ye, J. Wang, H. Yang, J. Yu, G. K.-S. Wong, and J. Wang, "ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun.," *PLoS Comput. Biol.*, vol. 1, no. 4, p. e43, Sep. 2005.
- [26] B. Rodríguez-Santiago and L. Armengol, "Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal," *Diagnóstico Prenat.*, vol. 23, no. 2, pp. 56–66, Apr. 2012.
- [27] M. Imelfort and D. Edwards, "De novo sequencing of plant genomes using second-generation technologies.," *Brief. Bioinform.*, vol. 10, no. 6, pp. 609–18, Nov. 2009.
- [28] R. Batista, "Las tres generaciones de la secuenciación," *¿ CÓMO FUNCIONA?*, vol. 3, no. 128, pp. 27–28, 2010.
- [29] J. L. García, "Aceleración de la Secuenciación Genómica (Rebajas en la secuenciación)," InstitutoRoche, 2009. [Online]. Available: http://institutoroche.es/bioinformatica/45/Aceleracion_de_la_Secuenciacion_Genomica_Rebajas_en_la_secuenciacion. [Accessed: 30-Dec-2014].
- [30] A. Rodríguez-Ulloa and R. Rodríguez, "Bioinformatic tools for

- proteomic data analysis: An overview," *Biotechnol. Apl.*, vol. 25, pp. 312–319, 2008.
- [31] NCBI Resource Coordinators, "Índice de /genomes/genbank/plant/," 2014. [Online]. Available: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/>. [Accessed: 30-Dec-2014].
- [32] N. R. Coordinators, G. E. Om-, M. Viewer, and S. Read, "Database resources of the National Center for Biotechnology Information.," *Nucleic Acids Res.*, pp. 1–12, Nov. 2014.
- [33] E. Paux, P. Sourdille, I. Mackay, and C. Feuillet, "Sequence-based marker development in wheat: advances and applications to breeding.," *Biotechnol. Adv.*, vol. 30, no. 5, pp. 1071–88, Jan. 2012.
- [34] J. A. Cabezas, J. Ibáñez, D. Lijavetzky, D. Vélez, G. Bravo, V. Rodríguez, I. Carreño, A. M. Jermakow, J. Carreño, L. Ruiz-García, M. R. Thomas, and J. M. Martínez-Zapater, "A 48 SNP set for grapevine cultivar identification.," *BMC Plant Biol.*, vol. 11, no. 1, p. 153, Jan. 2011.
- [35] Federación Nacional de Cafeteros, "Proyecto genoma del café en Colombia, inversión que da y seguirá dando frutos," *AL GRANO*, 2014. [Online]. Available: http://www.federaciondecafeteros.org/algrano-fnc-es/index.php/comments/proyecto_genoma_del_cafe_en_colombia_inversion_que_da_y_segura_dando_fruto/. [Accessed: 20-Jan-2015].
- [36] G. Cadena, "Desarrollos Científicos De Cenicafé En La Última Década," *Rev. Acad. Colomb.*, vol. 29, pp. 89–99, 2005.
- [37] Federación Nacional de Cafeteros-Cenicafé -, "Informe Anual," 2011.
- [38] R. C. Muñoz, "Aplicaciones de las nuevas tecnologías de secuenciación," *¿ CÓMO FUNCIONA?*, vol. 5, pp. 138–139, 2012.
- [39] J. Soto and C. Lopez, "RNA-seq : herramienta transcriptómica útil para el estudio de interacciones planta-patógeno," *Fitosanidas*, vol. 16, no. 2, pp. 101–113, 2012.
- [40] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage: key considerations in genomic analyses.," *Nat. Rev. Genet.*, vol. 15, no. 2, pp. 121–32, Feb. 2014.
- [41] K.-O. Mutz, A. Heilkenbrinker, M. Lönne, J.-G. Walter, and F. Stahl, "Transcriptome analysis using next-generation sequencing.," *Curr. Opin. Biotechnol.*, vol. 24, no. 1, pp. 22–30, Feb. 2013.
- [42] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics.," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, 2009.
- [43] A. Eduardo, R. Cubillos, L. P. Jiménez, M. Sc, A. Jimena, B. Giraldo, and D. Ph, "UNA REVISIÓN PARA NO EXPERTOS RNA-Seq Data Analysis in Prokaryotes : A Review for Non-experts," vol. 19, no. 2, pp. 131–142, 2014.
- [44] S. R. Strickler, A. Bombarely, and L. a Mueller, "Designing a transcriptome next-generation sequencing project for a nonmodel plant species.," *Am. J. Bot.*, vol. 99, no. 2, pp. 257–66, Feb. 2012.
- [45] E. R. Mardis, "A decade's perspective on DNA sequencing technology.," *Nature*, vol. 470, no. 7333, pp. 198–203, Feb. 2011.
- [46] P. L. Auer and R. W. Doerge, "Statistical design and analysis of RNA sequencing data.," *Genetics*, vol. 185, no. 2, pp. 405–16, Jun. 2010.
- [47] S. Anders and W. Huber, "Differential expression analysis for sequence count data.," *Genome Biol.*, vol. 11, no. 10, p. R106, Jan. 2010.
- [48] C. Sonesson and M. Delorenzi, "A comparison of methods for

differential expression analysis of RNA-seq data.," *BMC Bioinformatics*, vol. 14, no. 1, p. 91, Jan. 2013.

[49] X. Ruan and Y. Ruan, "Genome Wide Full-Length Transcript Analysis Using 5' and 3' Paired-End-Tag Next Generation Sequencing (RNA-PET)," in *Transcriptional Regulation*, Ales Vancu., J. Walker, Ed. 2012, pp. 535–562.

[50] M. J. Fullwood, C.-L. Wei, E. T. Liu, and Y. Ruan, "Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses.," *Genome Res.*, vol. 19, no. 4, pp. 521–32, Apr. 2009.

[51] ILLUMINA, "Single-Read Sequencing." [Online]. Available: <http://www.illumina.com/technology/next-generation-sequencing/single-read-sequencing-assay.html>. [Accessed: 14-Jan-2015].

[52] F. Ozsolak and P. M. Milos, "RNA sequencing: advances, challenges and opportunities.," *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 87–98, Feb. 2011.

[53] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang, "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.," *Gigascience*, vol. 1, no. 1, p. 18, Jan. 2012.

[54] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores.," *Genome Res.*, vol. 18, no. 11, pp. 1851–8, Nov. 2008.

[55] A. D. Smith, Z. Xuan, and M. Q. Zhang, "Using quality scores and longer reads improves accuracy of Solexa read mapping.," *BMC Bioinformatics*, vol. 9, no. 1, p. 128, Jan. 2008.

[56] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome.," *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–52, Jul. 2011.

[57] X. Zhou, H. Lindsay, and M. D. Robinson, "Robustly detecting differential expression in RNA sequencing data using observation weights.," *Nucleic Acids Res.*, vol. 42, no. 11, p. e91, Jan. 2014.

[58] C. Trapnell, B. a Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.," *Nat. Biotechnol.*, vol. 28, no. 5, pp. 511–5, May 2010.

[59] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq.," *Nat. Methods*, vol. 8, no. 6, pp. 469–77, Jun. 2011.

[60] I. Pagani, K. Liolios, J. Jansson, I.-M. A. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides, "The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D571–9, Jan. 2012.

[61] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq.," *Nat. Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[62] D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent,

"Expression profiling using cDNA microarrays," vol. 21, no. January, 1999.

[63] I. Nookaew, M. Papini, N. Pornputtpong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen, "A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*," *Nucleic Acids Res.*, vol. 40, no. 20, pp. 10084–97, Nov. 2012.

[64] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigó, "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression," *Genome Res.*, vol. 22, no. 9, pp. 1775–89, Sep. 2012.

[65] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn, "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses," *Genes Dev.*, vol. 25, no. 18, pp. 1915–27, Sep. 2011.

[66] R. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones, and M. Marra, "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing," *Biotechniques*, vol. 45, no. 1, pp. 81–94, Jul. 2008.

[67] R. Cullum, O. Alder, and P. a Hoodless, "The next generation: using new sequencing technologies to analyse gene regulation," *Respirology*, vol. 16, no. 2, pp. 210–22, Feb. 2011.

[68] J. Xu, Y. Li, X. Ma, J. Ding, K. Wang, S. Wang, Y. Tian, H. Zhang, and X.-G. Zhu, "Whole transcriptome analysis using next-generation sequencing of model species *Setaria viridis* to support C4 photosynthesis research," *Plant Mol. Biol.*, vol. 83, no. 1–2, pp. 77–87, Sep. 2013.

[69] X.-J. Han, Y.-D. Wang, Y.-C. Chen, L.-Y. Lin, and Q.-K. Wu, "Transcriptome sequencing and expression analysis of terpenoid biosynthesis genes in *Litsea cubeba*," *PLoS One*, vol. 8, no. 10, p. e76890, Jan. 2013.

[70] J. Gao, Y. Zhang, C. Zhang, F. Qi, X. Li, S. Mu, and Z. Peng, "Characterization of the floral transcriptome of Moso bamboo (*Phyllostachys edulis*) at different flowering developmental stages by transcriptome sequencing and RNA-seq analysis," *PLoS One*, vol. 9, no. 6, p. e98910, Jan. 2014.

[71] A. a Gullede, A. D. Roberts, H. Vora, K. Patel, and A. E. Loraine, "Mining *Arabidopsis thaliana* RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a," *Am. J. Bot.*, vol. 99, no. 2, pp. 219–31, Feb. 2012.

[72] X. Gan, O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Rättsch, and R. Mott, "Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*," *Nature*, vol. 477, no. 7365, pp. 419–23, Sep. 2011.

[73] K. I. L. H. Y. U. N. Kim, Y. A. N. G. J. A. E. Kang, D. O. N. G. H.

- Y. U. N. Kim, M. I. N. Y. O. Yoon, J. U. N. G. Y. Moon, M. O. O. N. Y. O. Kim, K. Y. Van, and S. U. K. A. Lee, "RNA-Seq Analysis of a Soybean Near-Isogenic Line Carrying Bacterial Leaf Pustule-Resistant and -Susceptible Alleles," *DNA Res.*, vol. 18, no. October, pp. 483–497, 2011.
- [74] L. Xu, L. Zhu, L. Tu, L. Liu, D. Yuan, L. Jin, L. Long, and X. Zhang, "Lignin metabolism has a central role in the resistance of cotton to the wilt fungus *Verticillium dahliae* as revealed by RNA-Seq-dependent transcriptional analysis and histochemistry.," *J. Exp. Bot.*, vol. 62, no. 15, pp. 5607–21, Nov. 2011.
- [75] G. E. Montoya, O. M. A. Cristancho, and A. M. D. P. M. Botero, "ANÁLISIS DE SECUENCIAS DE GENES DE *Coffea arabica* VAR. CATURRA," *Investig. Cenicafe*, vol. 57, no. 2, pp. 79–87, 2006.
- [76] University of Alberta, "1000 Plants." [Online]. Available: <https://sites.google.com/a/ualberta.ca/onekp/home>. [Accessed: 16-Dec-2014].
- [77] N. Matasci, L.-H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, S. Ayyampalayam, M. Barker, J. Burleigh, M. a Gitzendanner, E. Wafula, J. P. Der, C. W. dePamphilis, B. Roure, H. Philippe, B. R. Ruhfel, N. W. Miles, S. W. Graham, S. Mathews, B. Surek, M. Melkonian, D. E. Soltis, P. S. Soltis, C. Rothfels, L. Pokorny, J. a Shaw, L. DeGironimo, D. W. Stevenson, J. Villarreal, T. Chen, T. M. Kutchan, M. Rolf, R. S. Baucom, M. K. Deyholos, R. Samudrala, Z. Tian, X. Wu, X. Sun, Y. Zhang, J. Wang, J. Leebens-Mack, and G. K.-S. Wong, "Data access for the 1,000 Plants (1KP) project," *Gigascience*, vol. 3, no. 1, p. 17, 2014.
- [78] E. Asamizu, H. Ichihara, A. Nakaya, Y. Nakamura, H. Hirakawa, T. Ishii, T. Tamura, K. Fukami-Kobayashi, Y. Nakajima, and S. Tabata, "Plant Genome DataBase Japan (PGDBj): a portal website for the integration of plant genome-related databases.," *Plant Cell Physiol.*, vol. 55, no. 1, p. e8, Jan. 2014.
- [79] "Plant Program - DOE Joint Genome Institute." [Online]. Available: <http://jgi.doe.gov/our-science/science-programs/plant-genomics/>. [Accessed: 27-Jan-2015].
- [80] "Phytozome v10.1: Home." [Online]. Available: <http://phytozome.jgi.doe.gov/pz/portal.html>. [Accessed: 27-Jan-2015].
- [81] "Plant Flagship Genomes - DOE Joint Genome Institute." [Online]. Available: <http://jgi.doe.gov/our-science/science-programs/plant-genomics/plant-flagship-genomes/>. [Accessed: 27-Jan-2015].
- [82] "DOE JGI Plant Flagship Gene Atlas - DOE Joint Genome Institute." [Online]. Available: <http://jgi.doe.gov/doe-jgi-plant-flagship-gene-atlas/>. [Accessed: 27-Jan-2015].
- [83] S. A. Goff, M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, A. Muir, N. Merchant, S. Lowry, S. Mock, M. Helmke, A. Kubach, M. Narro, N. Hopkins, D. Micklos, U. Hilgert, M. Gonzales, C. Jordan, E. Skidmore, R. Dooley, J. Cazes, R. McLay, Z. Lu, S. Pasternak, L. Koesterke, W. H. Piel, R. Grene, C. Noutsos, K. Gendler, X. Feng, C. Tang, M. Lent, S.-J. Kim, K. Kvilekval, B. S. Manjunath, V. Tannen, A. Stamatakis, M. Sanderson, S. M. Welch, K. A. Cranston, P. Soltis, D. Soltis, B. O'Meara, C. Ane, T. Brutnell, D. J. Kleibenstein, J. W. White, J. Leebens-Mack, M. J. Donoghue, E. P. Spalding, T. J. Vision, C. R. Myers, D. Lowenthal, B. J. Enquist, B. Boyle, A. Akoglu, G. Andrews, S. Ram, D. Ware, L. Stein, and D. Stanzone, "The iPlant Collaborative: Cyberinfrastructure for Plant Biology.," *Front. Plant Sci.*, vol. 2, p. 34, Jan. 2011.

- [84] "The Project | iPlant Collaborative." [Online]. Available: <http://www.iplantcollaborative.org/about-iplant/the-project>. [Accessed: 27-Jan-2015].
- [85] "1001genomes.org." [Online]. Available: <http://1001genomes.org/>. [Accessed: 27-Jan-2015].
- [86] P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, and E. Huala, "The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1202–10, Jan. 2012.
- [87] "TAIR - Home Page." [Online]. Available: <http://www.arabidopsis.org/>. [Accessed: 17-Dec-2014].
- [88] N. Fernandez-Pozo, N. Menda, J. D. Edwards, S. Saha, I. Y. Teclé, S. R. Strickler, A. Bombarely, T. Fisher-york, A. Pujar, H. Foerster, A. Yan, and L. a Mueller, "The Sol Genomics Network (SGN)- from genotype to phenotype to breeding," *Nucleic Acids Res.*, no. 8, pp. 1–6, 2014.
- [89] "Sol Genomics Network." [Online]. Available: <http://solgenomics.net/>. [Accessed: 27-Jan-2015].
- [90] J. Duvick, A. Fu, U. Muppirala, M. Sabharwal, M. D. Wilkerson, C. J. Lawrence, C. Lushbough, and V. Brendel, "PlantGDB: A resource for comparative plant genomics," *Nucleic Acids Res.*, vol. 36, no. December 2007, pp. 959–965, 2008.
- [91] "PlantGDB - Resources for Plant Comparative Genomics." [Online]. Available: <http://www.plantgdb.org/>. [Accessed: 27-Jan-2015].

INICIATIVAS DE FORMACIÓN RELACIONADAS CON BIOINFORMÁTICA Y/O BIOLOGÍA COMPUTACIONAL EN COLOMBIA

En esta instancia se presentaron algunas iniciativas de formación académica con que cuenta el país en términos de bioinformática y biología computacional. A continuación se presentan los resultados de un estudio elaborado por BIOS en torno a la oferta y demanda de formación en temas bioinformáticos.

POSGRADOS EN EL ÁREA

Maestría en bioinformática y biología computacional de la red Alianza Suma de Manizales

La iniciativa de crear el programa fue concebida por la Alianza Suma en el año 2011. El objetivo del programa es “formar magísteres en Biología Computacional y Bioinformática con la finalidad de contribuir al fortalecimiento de personal calificado para modelar, sistematizar, simular procesos, interpretar y analizar información biológica obtenida a partir de la biodiversidad con aplicación en los sectores médico, industrial, farmacéutico, cosmético y agropecuario para generar productos, bienes y servicios que apoyen la solución de problemas de la región y el país”. Cuatro universidades de la ciudad se han asociado para ofrecer este programa de posgrado, que iniciará actividades durante el primer semestre del año 2015. Tales instituciones son: Universidad de Caldas, Universidad Autónoma de Manizales, Universidad Católica de Manizales y Universidad de Manizales. El programa cuenta con el apoyo de CENICAFÉ, BIOS y de la Fundación Luker.

El programa incluirá 60 créditos, distribuidos en 4 semestres, y acogerá 20 estudiantes por cohorte, distribuidos equitativamente entre las 4 universidades participantes.

Las líneas de investigación de la Maestría serán:

- Análisis y desarrollo de software para big data en biología.
- Modelado computacional de sistemas biológicos.
- Salud y genómica computacional.
- Métodos computacionales para el estudio de la biodiversidad y los biorrecursos.

Maestría en biología computacional de la Universidad de los Andes

El objetivo de la maestría es “formar profesionales altamente calificados, con capacidad de desarrollar investigación pura y aplicada en distintas líneas de investigación relacionadas, entre otros, con el análisis computacional de macromoléculas, modelamiento y simulación de sistemas biológicos y técnicas computacionales, además de contribuir al mejoramiento de la docencia en biología computacional y bioinformática, ciencias biológicas, física, matemáticas e ingeniería entre otras disciplinas en el país, formando profesionales”.

Cuenta con 40 créditos, distribuidos en 4 semestres académicos, con la posibilidad de elegir entre dos modalidades, investigación o profundización. Dentro del programa participan los departamentos de Ciencias Biológicas, Física, Ingeniería Civil y Ambiental, Ingeniería Química e Ingeniería de Sistemas y Computación.

El posgrado está enfocado a tres áreas o perfiles:

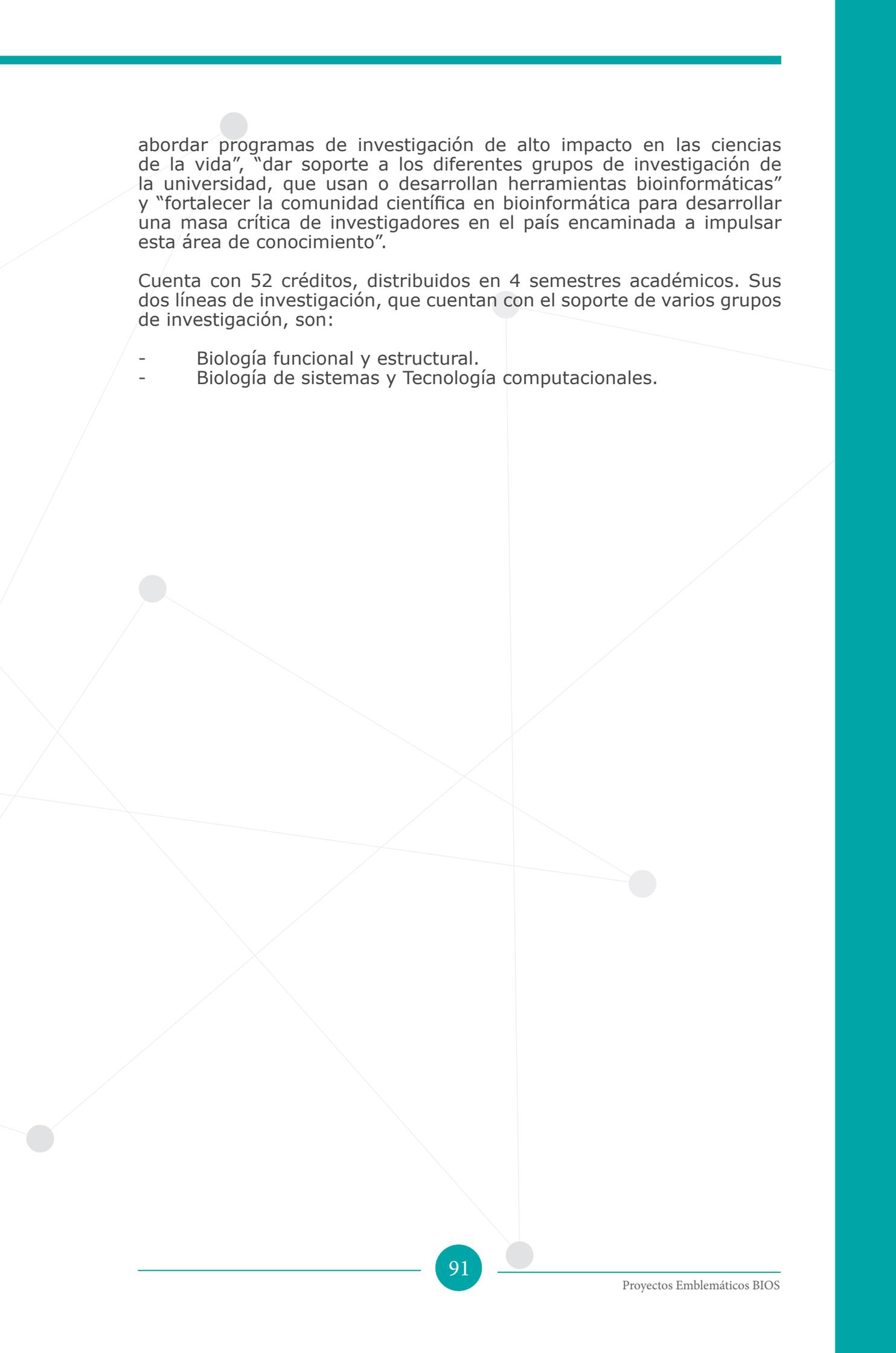
- Análisis de macromoléculas (análisis de secuencias biológicas, filogenética, predicción de estructuras tridimensionales, entre otros).
- Modelamiento y simulación de sistemas biológicos (modelamiento matemático de procesos biológicos a nivel de sistemas complejos completos).
- Técnicas computacionales (desarrollo de estrategias computacionales y su implementación para el análisis de problemas biológicos).

Maestría en biología computacional de la Universidad de los Andes

Se trata de un “programa de posgrado interdisciplinario dirigido a la formación avanzada de profesionales en bioinformática capaces de plantear, analizar y resolver problemas en ciencias de la vida, fortalecer este campo de conocimiento y consolidar una comunidad científica que contribuya al avance del conocimiento biológico”, con dos perfiles profesionales: investigación y profundización.

- Modalidad de profundización: sus objetivos son “formar profesionales que integren conocimiento en ciencias biológicas e ingeniería para diseñar proyectos encaminados a resolver problemas en ciencias de la vida”, “capacitar profesionales en el uso de técnicas y metodologías propias de la bioinformática para organizar y analizar datos de interés biológico”, “generar grupos interdisciplinarios capaces de abordar proyectos de alto impacto en las ciencias de la vida” y “dar soporte a los diferentes grupos de investigación de la universidad, que usan o desarrollan herramientas bioinformáticas”.

- Modalidad investigación: sus objetivos son “formar investigadores que integren conocimiento en ciencias biológicas e ingeniería para diseñar proyectos de investigación encaminados a resolver problemas en ciencias de la vida”, “capacitar investigadores en el uso de técnicas y metodologías propias de la bioinformática para organizar y analizar datos de interés biológico”, “generar grupos interdisciplinarios para



abordar programas de investigación de alto impacto en las ciencias de la vida”, “dar soporte a los diferentes grupos de investigación de la universidad, que usan o desarrollan herramientas bioinformáticas” y “fortalecer la comunidad científica en bioinformática para desarrollar una masa crítica de investigadores en el país encaminada a impulsar esta área de conocimiento”.

Cuenta con 52 créditos, distribuidos en 4 semestres académicos. Sus dos líneas de investigación, que cuentan con el soporte de varios grupos de investigación, son:

- Biología funcional y estructural.
- Biología de sistemas y Tecnología computacionales.

CONSIDERACIONES FINALES DEL “ENCUENTRO PARA LA IDENTIFICACIÓN DE PROYECTOS EMBLEMÁTICOS Y CONFORMACIÓN DE UNA COMUNIDAD BIOINFORMÁTICA”

- Las diferentes entidades participantes han manifestado su ánimo de hacer parte de una Comunidad Bioinformática y cada una de ellas cuenta con diferentes potenciales y áreas de acción que, al integrarse, lograrían consolidar una unidad fortalecida.
- Serían dos los proyectos emblemáticos a desarrollar en el seno de la Comunidad Bioinformática: HapMap Colombia y 1.000 transcriptomas.
- El ámbito de la bioinformática y la biología computacional es promisorio en Colombia y su incentivo representaría significativos beneficios para diferentes sectores del país.
- Si bien a nivel nacional se cuenta con el recurso humano, la capacidad de cómputo y la infraestructura necesaria para llevar a cabo ciertas investigaciones, aún falta articulación y coordinación interinstitucional, de manera que se evite la redundancia de esfuerzos y se fortalezcan proyectos de alto impacto.
- Los obstáculos burocráticos existentes en torno al acceso a los recursos genéticos representan una verdadera limitación para la puesta en marcha de diferentes iniciativas de investigación, razón por la cual, generar cambios a nivel normativo será vital para la promoción del sector bioinformático en el país.
- BIOS podría consolidarse como un eje articulador entre la academia, la empresa y el Estado, y para fortalecerse como centro de investigación podría poner en consideración la creación de un comité científico conformado por profesionales de diferentes partes del país.
- A la hora de emprender cualquier iniciativa de impacto nacional habrá de tenerse en cuenta tanto el valor social, como el valor de conocimiento y el valor de uso o impacto del proyecto.
- BIOS podría generar una línea de comunicación periódica, es decir, una publicación en la que se presenten detalles sobre avances relacionados con bioinformática y biología computacional. Para ello podría hacer uso de su página web o considerar la posibilidad de poner en circulación un boletín informativo.

- Es innegable el valor de la interdisciplinariedad a la hora de fortalecer cierta área de acción. En ese sentido, sería fundamental consolidar una estrategia nacional de I+D+I.
- Para próximos encuentros sería importante ampliar la participación del sector empresarial y para ello una estrategia eficaz sería extender la convocatoria a las Cámaras de Comercio. También habrán de citarse otros entes gubernamentales como los Ministerios de Medio Ambiente, Salud y Tecnologías de la Información y las Comunicaciones.

Una vez finalizado el "Encuentro para la Identificación de Proyectos Emblemáticos y Conformación de una Comunidad Bioinformática", los participantes acordaron:

- Intercambiar información para identificar oportunidades de investigación y desarrollo para las entidades que representan.
- Realizar reuniones periódicas para establecer estrategias de fortalecimiento del sector bioinformático del país. El primer encuentro tendría lugar durante el próximo mes de marzo de 2015 en la ciudad de Bogotá (los representantes de UNIANDES y UNAL gestionarán la posibilidad de realizarlo en sus respectivas instituciones). Una reunión posterior se realizaría en el mes de junio en el marco del curso de HPC (que se realizará en las instalaciones de BIOS). Otro escenario de encuentro sería el Congreso de Biología Computacional, el cual tendrá lugar en el mes de septiembre (Andrés Mauricio Gómez, de la Universidad de Antioquia, se ofreció a gestionar el espacio para el encuentro y para ello requiere un oficio de solicitud formal).
- En el próximo encuentro se realizarán postulaciones para liderar la Comunidad de Bioinformática en consolidación.
- Participar en los dos proyectos emblemáticos definidos para el país, los cuales tendrían el potencial de propiciar el trabajo articulado entre las diferentes entidades. Tales iniciativas son: HapMap.co y 1000 transcriptomas.
- Plantear la debida justificación y marco teórico de los proyectos emblemáticos definidos. BIOS realizará un esbozo inicial del texto y para ello será vital contar con el asesoramiento y trabajo participativo de las diferentes instituciones.
- Propiciar el desarrollo adecuado de los proyectos emblemáticos, para lo cual sería relevante contar con el asesoramiento de investigadores que tienen experiencia en dichas iniciativas a nivel internacional.
- Dialogar en torno a la verdadera viabilidad y utilidad de los proyectos, de manera que éstos no sean réplicas de iniciativas ya existentes y que puedan tener impacto a diferentes niveles (social, económico, científico, entre otros).
- Propiciar la articulación interinstitucional, para lo cual será fundamental:
 - Recopilar información sobre los grupos de investigación existentes

y crear un espacio donde dichos datos puedan ser publicados y puestos al alcance de toda la comunidad.

- Crear una página web de la comunidad.
- Integrarse a una plataforma habilitante en bioinformática con centros que trabajen en diferentes sectores y aumenten sus posibilidades de acceder a recursos.
- Generar un plan de trabajo estratégico entre las instituciones vinculadas a la comunidad para evitar redundancia de esfuerzos.
- Facilitar la interacción con el sector empresarial e industrial para detectar cuáles son sus principales necesidades en términos de herramientas bioinformáticas.
- Establecer un networking o un posible workshop para mantener la contextualización en temas bioinformáticos a nivel global.
- Abordar los siguientes aspectos en próximos encuentros:
 - Retomar las consideraciones recopiladas en la presente relatoría para definir las directrices, el plan de acción y los responsables de la Comunidad.
 - Entablar diálogo sobre la priorización de los subproyectos que integrarían los dos Proyectos Emblemáticos acordados.
 - Establecimiento de un marco formal que permita realizar las acciones de la Comunidad conformada bajo un escenario normativo y legal claro.
 - Evaluar y discutir los avances del estudio de oferta y demanda académica.
 - Crear un frente de acción con el potencial de inducir cambios reales en el marco normativo de acceso a los recursos genéticos.

EL FUTURO DE LA BIOLOGÍA COMPUTACIONAL Y LA BIOINFORMÁTICA EN LATINOAMÉRICA: UN RETO PARA FAVORECER LA SALUD, EL AGRO Y LA BIODIVERSIDAD

BIOS: HISTORIA, PROYECTOS Y RETOS DE UN CENTRO DE BIOINFORMÁTICA EN COLOMBIA

The background features a complex geometric design. On the left, there is a dense cluster of overlapping triangles in various shades of blue and teal. A thick, white diagonal line runs from the top left towards the bottom right, bisecting the composition. To the right of this line, a solid teal rectangular area contains the main text. The overall design is modern and technical, with faint grey lines and circular nodes scattered across the white background, suggesting a network or data structure.

ARTÍCULOS DE INTERÉS



EL FUTURO DE LA BIOLOGÍA COMPUTACIONAL Y LA BIOINFORMÁTICA EN LATINOAMÉRICA: UN RETO PARA FAVORECER LA SALUD, EL AGRO Y LA BIODIVERSIDAD



Mauricio Corredor Ph.D

Grupo GEBIOMIC, Instituto de Biología,
Facultad de Ciencias Exactas y Naturales ,
Universidad de Antioquia

INTRODUCCIÓN

Nuevos descubrimientos nos dejaron recientemente sorprendidos: un nuevo código genético oculto en vía a ser dilucidado (1, 2) y algunos ncRNAs que presentan traducción sin estar en un aparente ORF (3). Años atrás los descubrimientos sobre todas las enormes variantes de RNA y sus metilaciones nos confirman que será necesario mucha más información e investigación para dilucidar la verdadera relación entre lo digital y analógico de la información genética, epigenética y molecular dentro de las células. Adicional a ello los avances tecnológicos nos dejan perplejos sobre aparatos y equipos para desentrañar, genomas y proteomas completos a velocidades enormes. Existen equipos que ya realizan el secuenciamiento de un genoma completo en cuestión de horas, incluida la anotación o proteomas, que ya se podrán obtener en una semana.

Estamos frente a la acumulación de información jamás vista por la especie humana. Servidores que hablarán todos los idiomas almacenados en teraflops y petaflops (entendiendo flop como sigla del inglés para Floating-point Operations Per Second). Ya no hablamos de cantidad de información (en bits) sino más bien de performance o eficiencia para analizar esos yottabits o mejor Yottabytes (1 byte son 8 bits o sea un yottabyte es 1024 bytes). El Big Data es el ahora y será una historia contada ya en cuestión de unos cuantos años, pues ya sabemos cómo producir mucha información. El problema será como analizarla lo mejor y más eficientemente posible. Ante estos avances, los países desarrollados tienen la sartén por el mango en cuanto el almacenamiento de la información: en el siglo XX eran propietarios del 95% de las principales revistas científicas, y en el siglo XXI, dueños de un porcentaje parecido en el almacenamiento de la información. Pero en el segundo caso mucha de esa información es pública. ¿Cuál será la esperanza científica para los países que aún no alcanzamos el mundo desarrollado, al menos en bioinformática o en biocomputación?

Recordemos que con el surgimiento de la bioinformática cambió completamente la forma de hacer biología. En esas transformaciones contribuyeron especialmente la inserción de la química computacional, la ingeniería de software y los avances en matemáticas y estadística computacionales, haciendo de la bioinformática un abordaje completamente diferente para estudiar la vida, especialmente el mantenimiento de la vida humana. La industria farmacéutica y la biotecnología no pueden prescindir hoy día de los avances en genómica y proteómica para generar una medicina más personalizada; no obstante, el problema es que no todas las compañías compartirán sus descubrimientos en moléculas y software (4).

EL RETO PARA NUESTROS PAÍSES

En América Latina, una región mega-biodiversa, aún no sabemos cómo empoderarnos de toda nuestra información biológica, a veces sin legislación clara para algunos países o ahogados en cientos de artículos de reglamentación para otros; sin herramientas científicas a veces y sin presupuesto suficiente para invertir en protección, salud y prevención, entre otros. Ante ello, tenemos un papel crucial: aprovechar el conocimiento actual para los años por venir con el propósito de participar en la carrera de los avances tecnológicos (cerrar la brecha) y de apropiarnos de toda la información de nuestras reservas biológicas (sin excepción la humana) para no seguir perdiéndola de forma irrisoria sin hacer nada. O nos responsabilizamos de toda la información de nuestra biodiversidad o dejaremos pasar nuevamente un momento crucial en la historia de la ciencia. Ese reto de nuestros países latinoamericanos, en vías reales de desarrollo en la biología computacional y la bioinformática, ya es una realidad en países no desarrollados como Indonesia, Malasia, Tailandia e India. Los cuatro, países tropicales. En nuestro continente el tema lo lidera Brasil. ¿Y los restantes países? ¿Será que la bioinformática puede contribuir en el desarrollo o es una aparente ilusión tecnológica?

En ciencia la cuestión no es de creer o no. Los resultados son la base del trabajo futuro. La bioinformática y la biología computacional nos propusieron avances sorprendentes sobre las mejoras de nuestra salud y eso no fue una promesa. Un simple ejemplo de esos resultados es el del programa Inglés del desarrollo de 100.000 genomas (5): la personalización diagnóstica y la personalización terapéutica. Esos avances ya podrán ser utilizados en unos cuantos años, y muchos pacientes en este país serán tratados en un futuro no muy lejano sobre esa base tecnológica.

A nadie le cabe duda hoy en día que la bioinformática va más rápido que la biotecnología, y es que ha sido precisamente la bioinformática la que le dio un desarrollo más veloz a la biotecnología. Lo que quiere decir que con estas dos disciplinas unidas se puede lograr desarrollos tecnológicos y resultados sorprendentes para las mejoras en nuestra salud, bienestar, calidad del aire, el agua y el suelo. De otra parte la nanotecnología propone alternativas importantes en terapéutica haciendo blanco con objetivos moleculares sobre tejidos específicos, abandonando el concepto de regadera en los tratamientos, y permitiendo así ahorrar insumos, gastos en tratamiento, tiempo en resultados y finalmente lo más importante: salvaguardar más vidas. El ejemplo más importante lo tiene en sus manos la industria farmacéutica, con la denominada fármaco-genómica que marcará el avenir de la fabricación de nuevos medicamentos muy distintos a los que todavía consumimos hoy. Si aproximadamente 25.000 genes del ser humano no se expresan igual en cada individuo, no tiene sentido consumir sustancias en tratamientos estandarizados entre 2 y 10 modelos simples para 7.000 millones de personas.

Surge entonces la pregunta sobre el futuro de nuestro subcontinente: ¿están preparadas las industrias farmacéuticas latinoamericanas para estos desarrollos tecnológicos? Si lo están, serán industrias líderes que competirán con las de los países desarrollados; de no ser así, su futuro

será incierto. Mejor aún, una medicina apropiada a las necesidades de nuestras poblaciones y genomas estará más adaptada a las industrias farmacéuticas nuestras. Sin embargo, recordemos los logros mundiales en la fabricación de vacunas, las cuales, en principio, no fueron desarrolladas por nuestros países. El ejemplo más claro fue el desarrollo de la vacuna de la fiebre amarilla, que hoy tenemos la capacidad de fabricar, pero que debemos importar desde países desarrollados hacia varios de nuestros países, pues fiebre amarilla no da en Europa o Norteamérica. Lo mismo sucede hoy con el Dengue, Chicunguña o Ébola, que genera muertos en nuestras regiones, pero los desarrollos tecnológicos para su prevención siguen siendo importados. Parece ser que la regla para nuestros países sigue siendo: "es mejor importar tecnología que crearla; o, es mejor copiar o reproducir tecnología que inventarla".

GENÓMICA

Grandes bases de datos mundiales albergan miles de genomas, desde bacterias, como PATRIC (6), y de humanos, como 1000 Genomes (7). En la primera se encuentran varios genomas bacterianos de pacientes latinoamericanos, y en la segunda reposan ya genomas de latinoamericanos y hasta 60 genomas colombianos (de Antioquia). Sin embargo, en esos avances hay mucho trabajo por hacer y aunque en tales genomas la materia prima es nuestra, ellos no fueron ensamblados por nuestra propia tecnología. Más aún, son pocos los Latinoamericanos que ponen sus manos en investigar sobre esta información que es nuestra. Salvo algunas excepciones, la regla no ha sido la apropiación de nuestra biodiversidad usando la genómica, o ésta sigue siendo construida por quienes tienen la tecnología, por una simple y sencilla razón: inversión. Tenemos hoy el patrimonio científico, pero algunos de nuestros gobiernos aún creen que invertir en ciencia y tecnología de punta no solo es un capital de exagerado riesgo, sino quizás un capital a pérdida.

La industria Latinoamericana podría ser una excelente alternativa para la búsqueda de inversión, pero nuestros científicos e industriales marchan sobre planos tan diferentes: la genómica puede ser ciencia ficción para algunos inversionistas locales, pero para algunos de nuestros estudiantes en biología, bioingeniería y hasta ingeniería de sistemas construir y emplear un genomsa comienza a ser rutina. Por otro lado, es de conocimiento público que no solo inversionistas latinoamericanos invierten en universidades americanas, sino que ya lo hacen en proyectos de genómica internacionales. Sin embargo, hay ejemplos que están cambiando ese paradigma y no todo es desconocimiento en nuestros inversionistas Latinoamericanos, como el caso de la alianza BROAD-Instituto Nacional de Medicina Genómica de México. Esto puede ser incluso una experiencia de aprendizaje común para después revertir ese aprendizaje en nuestro patrimonio y en nuestra gente (8), especialmente en aspectos claves tales como los que se mencionan en uno de sus videos: "cerrar la brecha tecnológica" (9). El asunto hoy no es importar tecnología, es crearla, y en este sentido la computación y la genómica pueden ser un despertar que impulse nuestra investigación y nuestra ciencia. Algunos proyectos de genómica han fallado en nuestros países por la compra de tecnología sin tener aún la inversión garantizada para 5 o 10 años, como mínimo.

PROTEÓMICA

Brasil, Argentina, Chile y México ya cuentan con institutos o unidades exclusivas en proteómica, que la consideramos con transcriptómica, la gran segunda revolución de las ómicas. Incluso Brasil ya tiene la Sociedad Brasileira de Proteómica (10). Es decir, después de pasar por congresos nacionales de Bioinformática y Genómica, ya tiene congresos de proteómica (11), haciendo alarde ya no de un Brasil de la bioinformática o la biología computacional o la genómica, sino de un Brasil de la Proteómica como lo titula un artículo en Proteomics (12). El mismo caso sucede en México donde ya se cuenta con una Sociedad Mexicana de Proteómica (13). Ello demuestra que la apropiación científica del conocimiento de punta en Latinoamérica no es difícil cuando hay inversión para apropiación. Tenemos la capacidad de absorber conocimiento fácilmente y revertirlo en nuestros estudiantes y profesionales, sin embargo en este caso se recalca nuevamente que mientras más compleja sea la ciencia más difícil es hacerla entender a nuestras instituciones y entidades gubernamentales. Parece evidente que nuestros gobernantes tiene claro que es necesario hacer inversiones para educación de alto nivel como lo es en maestrías y doctorados nacionales o internacionales, pero no parece muy claro por qué en algunos países la inversión solo debe llegar hasta allí y no continuar incentivando la investigación. Nuevamente el ejemplo anterior de la genómica podría cambiar la mirada de los inversionistas latinoamericanos y las industrias farmacéuticas regionales pueden seguir el ejemplo de países desarrollados, que utilizan con éxito la biotecnología, el ADN recombinante y la transgénesis como es el caso nuestro de la farmacéutica BIOSIDUS en Argentina (14). Para la Asociación Latinoamericana de Industrias Farmacéuticas éste sería un gran reto, no solo en lo biológico sino también en lo computacional. Sin querer desmeritar el papel de esta asociación, al hacer una búsqueda en su página web las palabras genómica y proteómica aún no aparecen como tema. Bien es cierto que patentamos muy poco en la industria Latinoamericana por la falta de actualización de los modelos de investigación y desarrollo industrial.

TRANSCRIPTÓMICA Y EPIGENÓMICA

No es necesario crear para cada ómica un instituto o laboratorio, aunque ese sea el caso a seguir en algunos países desarrollados. Sin embargo, genómica, proteómica y transcriptómica son temas interrelacionados para muchas instituciones mundiales que no requieren independizar siempre. De hecho el mismo principio tecnológico y computacional se utiliza para obtener genomas y transcriptomas, por tanto, la integración digital (Next Generation Sequencing) con lo analógico (3D de RNA y proteínas) será crucial en la dilucidación de nuevos códigos a ambos niveles. La transcriptómica se ha desarrollado enormemente, dado que hoy sabemos de la existencia de una enorme cantidad de ncRNA (no codificantes) que regulan un gran conjunto de funciones. Pensábamos no hace más de 15 años que esas funciones estaban casi completamente reservadas a las proteínas; el asunto de los intrones sonaba más como algo peculiar que como un asunto clave. Y efectivamente hoy sabemos que existe toda una nueva genética molecular basada simplemente

en RNA, que las mutaciones no son solo un asunto de nucleótidos de DNA sino que subyace en él un nuevo tipo de mutación efectuada por metilaciones, acetilaciones (epigenética), entre otras, que comparte con el mismo RNA.

La epigenómica es quizás uno de los campos con más futuro para estudiar algunas enfermedades y la nutrición de la población Latinoamericana que ha padecido enormes dificultades. Pues bien, los primeros estudios en epigenética, demostraron que una hambruna en la población sueca de Overkalyx causó daños irreparables en las generaciones siguientes, padeciendo las consecuencias heredadas de sus ancestros (15, 16, 17). ¿Cómo será el caso de nuestras poblaciones Indoamericanas que han padecido más de 500 años de sufrimientos y vicisitudes? ¿Qué cambios epigenéticos específicos han padecido nuestras poblaciones? En la genética clásica los genes se pasan de padres a hijos sin sus experiencias, pero la epigenética nos demuestra que las experiencias de padres e hijos pasan de generación en generación. Tenemos entonces ya nuestro propio laboratorio genómico para estudiar la población, pues son nuestras propias experiencias las que determinaran nuestros genomas. Según este descubrimiento estamos frente a un problema genómico particular, una verdadera razón específica para entendernos y para saber por qué los Latinoamericanos somos como somos y no como deberíamos ser. Hay allí un motivo para abordar la medicina, la agronomía y la veterinaria como un asunto particular para generar nuestra propia información. Somos epigenéticamente las experiencias que vivimos, lo que comemos (nutri-epi-genómica) y lo que sentimos.

METABOLÓMICA, METEGENÓMICA Y OTRAS ÓMICAS

Estas áreas no solo se han desarrollado con componentes del metabolismo (legado de la bioquímica), como enzimas, sustratos y productos, sino igualmente dilucidadas también desde la formulación de ellas mismas, así como también por el desarrollo de otras nuevas ómicas como son la lipidómica, sacarómica y la generación de microbiomas con metagenómica. Por ejemplo para cuenta a donde hemos llegado en el caso del humano y los animales, no solo somos nuestros genes enfrentados a lo que comemos, sino que los microorganismos que portamos sobre nuestro cuerpo también limitarán o favorecerán nuestra nutrición y fenotipo. El microbioma es un tema nuevo que ha desarrollado nuevas áreas de la bioinformática y la biología computacional, especialmente perfeccionado hoy por la metagenómica, permitiendo identificar cientos de especies bacterianas no cultivables de ambientes incluso extremos, en el suelo, aire y agua. Nuestros países deben tener los microbiomas más grandes de la Tierra. Bacterias, protistas, virus y hongos aún están en proceso de ser repertoriados ya no a nivel de 16S o 18S, mejor aún, a nivel de su genoma y su proteoma. Queda mucho trabajo por lograr algo promisorio a este nivel, sin ser menos importante que las tres ómicas anteriores. Para sorpresa nuestra, instituciones como NASA están empleando al metagenómica para demostrar que hay microorganismos a varios kilómetros más allá de la atmosfera oxigenada (troposfera), donde pueden permanecer suspendidos días, meses y hasta años (18). La troposfera tropical tiene más kilómetros y seguro más biodiversidad que la de los países templados. Podríamos pensar que ya estas ciencias tan lejanas de la Tierra como la astrobiología, solo estarán reservadas

para unos pocos países privilegiados: ¿excusas? Por esa razón dos países privilegiados Latinoamericanos tienen dos de los mejores telescopios del mundo en Chile e inversiones en la Estación Espacial Internacional como Brasil. Sí, en este caso nuevamente podríamos decir que las inversiones de poca ganancia como son en astronomía e investigaciones espaciales estaba reservada para privilegiados: Australia, India y China que tienen músculo económico no están allí. Pero, Australia, India y China son países desarrollados en bioinformática y biología computacional. Necesitaremos mucha mano de obra en Latinoamérica para estudiar más metabolómica, lipidómica, metagenómica, interactómica, psicogenómica, neurobiómica, nutriómica, fluxómica, epigenómica, teratómica y tantas ómicas que faltan quizás por aparecer (19).

BIOLOGÍA DE SISTEMAS

Como su denominación lo establece la biología de sistemas elabora el modelamiento computacional y matemático de los sistemas biológicos. Por lo tanto, la integración de las ómicas por la biología de sistemas no se hacía esperar. Incluso antes de ser una realidad la genómica, los institutos de Biología de Sistemas ya eran un hecho en los países desarrollados, previendo el futuro. Igualmente es claro que la generación de demasiadas ómicas podría desintegrar el análisis, que es contrario al propósito de la biología de sistemas (integración). Hay todavía muy pocos institutos o centros de biología de sistemas en Latinoamérica a diferencias de los que ya hay en países no desarrollados como Irán, Turquía, Malasia o Pakistán, entre otros (20). Esta área al igual que las nuevas ómicas son nuevas áreas promisorias para nuestros países. Qué tenga conocimiento, solo Chile, México y Brasil tienen un centro o instituto de biología de sistemas que para el caso chileno combina con la nanotecnología (21), pero es de recalcar que también México ha sido un país líder en el estudio de sistemas complejos a nivel mundial, previamente incubado también, en buena medida en Latinoamérica por los chilenos Humberto Maturana y Francisco Varela; considerados hoy por algunos en las ciencias complejas, biólogos de sistemas notables a nivel mundial. Ello nuevamente es muestra que con la generación de ideas y nuevas teorías podemos ser pioneros en el mundo sin estar soñando.

MODELAMIENTO COMPUTACIONAL

Dejemos un poco de lado los aspectos informáticos de las células y entremos más en aspectos computacionales y matemáticos. Para el modelamiento matemático y computacional, como el de la biología de sistemas no se necesita muchos datos. Se requiere más imaginación, capacidad de deducción, creatividad y conocimiento para generar nuevas ideas (creo que en Latinoamérica tenemos las mismas capacidades para generar nuevas ideas científicas). Lo mismo sucede con la biología de sistemas y el modelamiento matemático para generar nuevos modelos sobre un computador. A veces ni siquiera es necesario crear herramientas computacionales. Eso hicieron los teóricos en física, química y matemática. No estaba ni siquiera desarrollada la computación para concebir el modelo del átomo o el enlace covalente.

Para el modelamiento computacional no se requiere tener mucha teoría en la cabeza, más que simplemente capacidad de intuición basada en conocimientos. En Latinoamérica sabemos correr cientos de programas (repetir) pero generalmente no somos capaces de ver en el fondo, cientos de ideas escondidas (crear). Lo primero que debemos cambiar es ese ridículo mito. Faltan entonces teóricos con conocimiento y libertad para pensar, única y exclusivamente. La cuestión es que si no le damos espacio a nuestros investigadores y medios para crear, ¿cómo queremos que ellos produzcan inventos y generen nuevas ideas?

Miles de profesores e investigadores Latinoamericanos están ahogados en oficios administrativos, académicos, logísticos, presupuestales entre otros, que a veces lo único que alcanzan a producir son artículos con resultados que trascienden exclusivamente a nivel local. Tenemos millones de computadores en Latinoamérica y la mayoría de la población los utiliza para enviar mensajes, chatear y leer noticias de farándula, eso incluido la población universitaria a todos los niveles, pero, salvo contadas excepciones, no generamos información a diario, no creamos programas o herramientas, no encontramos una asociación entre modelar computacionalmente un problema de laboratorio con una célula, un aspecto económico, unas elecciones, un asunto estelar o químico, donde una herramienta con Matlab, R, C/C++, Python, Perl, o el viejo Fortran pueden hacer exactamente lo mismo con la misma fórmula matemática. Sin embargo, tengo que reconocer que a este nivel existe muchísimo trabajo en Latinoamérica casi en solitario, de muchos temas de investigación donde profesores y alumnos tratan de resolver una enorme cantidad de asuntos, usando simplemente algún programa para simular problemas estadísticos, económicos, climáticos, biológicos, moleculares y atómicos. A pesar de ello, científicos e ingenieros de sistemas Latinoamericanos aún siguen ensimismados todavía en sus problemas particulares sin entenderse unos a otros. Pero hay casos excepcionales: traigo a colación el caso de un colega médico que trabajó primero en redes neuronales en C+ y hoy es un importante miembro de sistemas de un importante banco nacional. ¿Quién lo creería? Capacidad e ingenio simplemente, pero también otra prueba más de quien tiene el dinero puede interesarse por nuestras investigaciones.

CLOUD, BIG DATA Y REDES COMPUTACIONALES

Sin la ingeniería computacional no existirían la bioinformática y la biología computacional, ese papel es necesario reconocerlo, y es clave que aquellos que trabajamos con células, modelos biológicos, enfermedades y tratamientos, reconozcamos que le debemos parte de nuestro trabajo y resultados al mundo de la ingeniería computacional (software, hardware, network). Ese logro tecnológico nos permite llevar un aparato en el bolsillo con más memoria y velocidad que los computadores que teníamos en casa en la década de los años 90. Semejante desarrollo computacional permitiría un desarrollo en internet, exponencial (así fue para muchos casos), lo que en nuestros países latinoamericanos pudimos haber aprovechado a una escala informática en biocomputo. No lo fue así para todos los países. Sin embargo, algunas redes informáticas Latinoamericanas se formaron en los últimos años con excelente intensiones, pero todavía sin reales efectos sobre nuestro desarrollo, al menos en lo bioinformático o computacional (22, 23, 24). ¿Qué falta

nuevamente? Una real integración, ya que muchos esfuerzos han sido casi particulares y no globales. Por supuesto que la red mejor construida a este nivel es la Associação Brasileira de Bioinformática e Biologia Computacional (AB3C) (25), que es un caso aparte. Alternativamente surgen ideas particulares como la de Wikilife, un proyecto argentino que pretende recoger información de los pacientes a través de su celular (26). De todos modos todavía la producción de grandes bancos de información de los servidores Latinoamericanos queda reservada a otras disciplinas como la de los bancos, las instituciones gubernamentales, la justicia, los impuestos y algunas cuantas industrias específicas de servicios como la de los celulares. Un ejemplo es el Data Center Triara de Claro en Bogotá (27). Pero ni siquiera las universidades Latinoamericanas son en sí mismas grandes generadoras de bancos de información informático. No existe una red en vías de formación como un NCBI, un EMBL o un EBI Latinoamericano, donde un gran número de servidores se encuentren integrados copiando la información que otros servidores estén generando. Simplemente existen grandes universidades con buena capacidad para generar información, pero no muchas instituciones pueden hacerse cargo del almacenamiento a gran escala, por una simple razón: mantener la información cuesta y solo algunas universidades pueden responsabilizarse de sus propios datos. Los costos energéticos y de mantenimiento son demasiado altos para asumir además una carga externa que no represente ganancias. Este es quizás el campo de la bioinformática de menos desarrollado en nuestra región y si no llegamos a ese nivel seguiremos haciendo la biología computacional de laptop, del pequeño servidor o de servidor local.

CONCLUSIONES: EL MAÑANA ES AHORA

Seguramente se me quedó entre el tintero la formación de algún nuevo instituto o centro en ómicas o biología de sistemas o red nacional, Latinoamericana(o) de bioinformática o de algún tipo de servidor o herramienta de biocomputación en Latinoamérica. A ellos van mis excusas, especialmente a los grupos Latinoamericanos de biología computacional, y bioinformática: los verdaderos líderes no mencionados, responsables de construir a futuro una red nacional y Latinoamericana de biología computacional (la lista es larga). Sobre todos estos grupos y redes nacionales recae ese sueño de hacer realidad la bioinformática como herramienta de progreso tecnológico, para el bienestar y desarrollo. Es por lo tanto necesario sensibilizar a los gobiernos, industriales y colegas para hacer de esta ciencia activa un verdadero renacer de la ciencia Latinoamericana. Ya Brasil lo tiene claro presupuestalmente (28) y en muchas áreas México, Chile, Argentina, Costa Rica, Cuba (29, 30, 31, 32) y Colombia (33, 34). Pero los retos de hace 7 u 8 años no son los mismos de ahora. Sin embargo, antes que nada la formación debe seguir en aumento y sobretodo es necesario tener claro cuál es el verdadero objetivo de dictar un curso en bioinformática o en biología computacional (35). Yendo más allá después de formar estudiantes, hacer investigación y publicar, será igualmente importante generar alianzas y formar redes, especialmente computacionales. Por eso es importante saber cuál es el objetivo final de construir una asociación o red nacional en bioinformática y/o en biología computacional (36). Ese es el reto que sigue para nuestro país gracias a lo logrado en los CCBCOLs: I-2011, II-2013 y ahora éste en el III-2015. Una asociación nacional en este campo debe ser

proactiva, debe permitir el intercambio de ideas, presupuestos, temas y alianzas: y creo que eso lo han logrado en buena parte GeBiX, BIOS, y el CNSG. Faltaría entonces más integración a nivel computacional. Por eso es necesario seguir así, no parar el ritmo y llegar más lejos. Además, la invitación es a que empresas con músculo informático y presupuestal implantadas en Latinoamérica como para citar algunos ejemplos: Claro y Google, entre otras, se vinculen e inviertan en nuestras propuestas en bioinformática, en biología computacional, en biología de sistemas, en modelamiento y después en un futuro quizás en Big Data. Todo ello para que sea aplicado en salud, agro y biodiversidad; como ya lo viene haciendo Slim-México (8) o como lo podrían hacer Google Cloud Plataform, que ya invierte en genómica (37). Finamente ayudaría mucho también tener una revista Latinoamericana que soporte igualmente un congreso Latinoamericano en Biología Computacional y Bioinformática.

AGRADECIMIENTOS

Al III Congreso Colombiano de Biología Computacional y Bioinformática de Medellín, a la Universidad Antioquia en Medellín y a BIOS en Manizales

BIBLIOGRAFÍA Y CIBERGRAFÍA

1. Weatheritt, Robert J., and M. Madan Babu. "The hidden codes that shape protein evolution." *Science (New York, NY)* 342.6164 (2013): 1325
2. Stergachis, Andrew B., et al. "Exonic transcription factor binding directs codon choice and affects protein evolution." *Science* 342.6164 (2013): 1367-1372.
3. Pauli, Andrea, Eivind Valen, and Alexander F. Schier. "Identifying (non) coding RNAs and small peptides: Challenges and opportunities." *BioEssays* 37.1 (2015): 103-112.
4. Vihinen, M. "No more hidden solutions in bioinformatics." *Nature* 521.7552 (2015): 261.
5. <http://www.genomicsengland.co.uk/the-100000-genomes-project/>
6. Wattam, Alice R., et al. "PATRIC, the bacterial bioinformatics database and analysis resource." *Nucleic acids research* (2013): gkt1099.
7. 1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes." *Nature* 491.7422 (2012): 56-65.
8. <http://www.salud.carlosslim.org/iniciativa-slim-en-medicina-genomica/>
9. <https://www.youtube.com/watch?v=Dm1yVHnGzMY>
10. <http://www.brprot.org/>
11. <http://onlinelibrary.wiley.com/doi/10.1002/pmic.v12.17/issuetoc>
12. Martins de Souza, Daniel. "Brazil: the country of proteomics." *Proteomics* 12.17 (2012): 2599-2600.
13. <http://www.smp.org.mx/nsmp/index.php>
14. <http://www.biosidus.com.ar/>
15. Pembrey, Marcus E., et al. "Sex-specific, male-line transgenerational responses in humans." *European Journal of Human Genetics* 14.2 (2006): 159-166.
16. Pembrey, Marcus E. "Male-line transgenerational responses in

- humans." *Human Fertility* 13.4 (2010): 268-271.
17. Kaati, Gunnar, et al. "Transgenerational response to nutrition, early life circumstances and longevity." *European Journal of Human Genetics* 15.7 (2007): 784-790.
 18. DeLeon-Rodriguez, Natasha, et al. "Microbiome of the upper troposphere: Species composition and prevalence, effects of tropical storms, and atmospheric implications." *Proceedings of the National Academy of Sciences* 110.7 (2013): 2575-2580.
 19. http://omics.org/index.php/Main_Page
 20. https://en.wikipedia.org/wiki/List_of_systems_biology_research_groups
 21. <http://www.usm.cl/investigacion/centros/nanotecnologia-y-biologia-de-sistemas/>
 22. <http://www.ciatnews.cgiar.org/es/tag/bioinformatica/>
 23. <http://www.co.embnet.org/>
 24. <http://www.risc-project.eu/partners/>
 25. <http://ab3c.cebio.org/>
 26. <http://www.wikilife.org.ar>
 27. <http://www.claro.com.co/wps/portal/co/pc/corporaciones/data-center/computacion-y-servicios-profesionales-ti>
 28. Neshich, Goran. "Computational biology in Brazil." *PLoS Comput Biol* 3.10 (2007): e185.
 29. Bourne, Philip E., and Steven E. Brenner. "Developing computational biology." (2007): e157.
 30. Bassi, Sebastian, Virginia González, and Gustavo Parisi. "Computational biology in Argentina." *PLoS Comput Biol* 3.12 (2007): e257.
 31. Moreno, Edgardo, Bruno Lomonte, and José-María Gutiérrez. "Computational biology in Costa Rica: The role of a small country in the global context of bioinformatics." *PLoS Comput Biol* 4 (2008): e1000040.
 32. Pons, Tirso, Luis A. Montero, and Juan P. Febles. "Computational biology in Cuba: An opportunity to promote science in a developing country." (2007): e227
 33. Restrepo, S., Pinzón, A., Rodríguez-R, L. M., Sierra, R., Grajales, A. et al. 2009. Computational Biology in Colombia. *PLoS Comput Biol* 5 (10): e1000535. doi:10.1371/journal.pcbi.1000535.
 34. Benítez-Paez A., Cárdenas-Brito S. 2010. Bioinformática en Colombia: presente y futuro de la investigación biocomputacional. *Biomedica*. 30(2).
 35. Via A, De Las Rivas J, Attwood TK, Landsman D, Brazas MD, Leunissen JAM, et al. (2011) Ten Simple Rules for Developing a Short Bioinformatics Training Course. *PLoS Comput Biol* 7(10): e1002245. doi:10.1371/journal.pcbi.1002245
 36. Budd A, Corpas M, Brazas MD, Fuller JC, Goecks J, Mulder NJ, et al. (2015) A Quick Guide for Building a Successful Bioinformatics Community. *PLoS Comput Biol* 11(2): e1003972. doi:10.1371/journal.pcbi.1003972
 37. https://cloud.google.com/genomics/?utm_source=google&utm_medium=cpc&utm_campaign=2015-q2-cloud-latam-solutions-skws-freetrial



BIOS: HISTORIA, PROYECTOS Y RETOS DE UN CENTRO DE BIOINFORMÁTICA EN COLOMBIA



Marco Aurelio Cristancho
Director Científico BIOS



Dago Hernando Bedoya
Coordinador Bionegocios

La historia del Centro de Bioinformática y Biología Computacional de Colombia se remonta al año 2007, cuando el dueño de Microsoft, Bill Gates, visitó el país con la idea de apoyar la tecnología y la innovación nacional desde su entidad. Con esta directriz, directivas de instituciones como Colciencias y el MinTIC, iniciaron la formulación de un proyecto en el que al unir el esfuerzo del gobierno con la entidad privada, a través del soporte de la academia, se diera inicio a un centro de investigación de alto nivel en Colombia, que tuviera como meta aprovechar sostenible y sustentablemente la biodiversidad y su potencial para el desarrollo e impacto en nuevos sectores económicos latinoamericanos y mundiales.

Al realizar un exhaustivo estudio en el mundo sobre las tendencias científicas más significativas y trascendentales que estuvieran generando impactos positivos sobre los procesos económicos se encontró que la bioinformática y la biología computacional eran disciplinas que cumplían con dichas características, además de ser un apoyo importante para temas como biociencias, bioproductos, biotecnología, servicios ambientales, el mundo de las TICs, ciencias de la tierra, y salud, entre otras. Esta ha sido la respuesta del gobierno nacional (desde Colciencias y MinTIC) para la Ciencia, Tecnología e Innovación de alto nivel y con valor agregado, temas trascendentales actualmente en la economía mundial.

BIOS inició propiamente sus labores de servicio e investigación en el 2013 con la llegada del súper computador. La infraestructura computacional de este equipo está basada en servidores Hewlett-Packard (HP) con tecnología GP GPU (paraleliza, hace que un proceso que se ejecuta, sea optimizado) y coprocesadores INTEL - XEON PHI, esto quiere decir, que gracias a las características del equipo se pueden hacer trabajos de análisis de datos, almacenamiento, entre otros. Para el tema de análisis de datos en bioinformática, esta apuesta tecnológica en supercomputación es la más adecuada. Para almacenamiento contamos con una red SAN (Storage Area Network), con una capacidad de almacenamiento de 282 terabytes y para procesamiento tenemos cerca de 3.5 terabytes de RAM. Estamos al 10 por ciento de lo que realmente se quiere implementar en infraestructura computacional (Tabla 1 o Figura 1).

BIOS es un centro que está a la vanguardia mundial en herramientas de supercomputación y de visualización de datos biológicos. Desde el 2013 se inició la operación de un proyecto de regalías, patrocinado por el Fondo para la Ciencia, Tecnología e Innovación, a través de la Gobernación de Caldas, denominado Caldas Bioregión. Esta iniciativa tiene como meta convertir al departamento de Caldas en la región referencia del país en sinergia entre empresa, estado y academia en temas biotecnológicos.

Contamos también con el muro de visualización de datos científicos más importante del país. Este muro está compuesto por 32 pantallas de 55 pulgadas cada una, con una resolución individual de 1920x1080 pero que al combinarla logra la asombrosa suma de 66 millones de píxeles, 66 millones de puntos de luz para hacer ciencia. Esta herramienta servirá para visualizar la gran cantidad de datos que manejan BIOS y sus aliados, modelar moléculas y nuevos fármacos y el desarrollo de aplicativos para la agroindustria y el sector salud.

La misión del centro es proporcionar herramientas computacionales en relación con las ciencias de la vida, desarrollar software, almacenar y procesar datos suministrados por sus investigadores, por miembros de otros grupos de investigación, o por terceros; adelantar actividades académicas y educativas, asesorar y orientar investigaciones intensivas en el uso de datos y su interpretación y participar en el desarrollo de investigaciones, que requieran el uso de sus facilidades de procesamiento de datos en otras áreas del conocimiento, de importancia para el desarrollo de la ciencia, la tecnología y la innovación.

La visión de BIOS está dirigida a ser un centro de referencia en América Latina en la generación, procesamiento y aplicación de conocimiento en biodiversidad, para beneficio de empresas, universidades, centros y grupos de investigación, con los más altos estándares de calidad y confidencialidad.

TRANSFERENCIA TECNOLÓGICA

En BIOS hemos identificado la necesidad de estimular la oferta y demanda del sector biotecnológico, fortaleciendo la cultura Bio apoyada con TIC, mediante la difusión, la formación de alto nivel y la transferencia tecnológica. El Centro ofrece permanentemente cursos de corta duración en de análisis e interpretación de datos biológicos, visualización de anotación de genomas, estadística en el análisis de datos genómicos y diversas áreas de la biotecnología y bionegocios. BIOS cuenta además con un portafolio de 7 diplomados relacionados con las líneas de investigación del centro y participa como aliado en la maestría en bioinformática y biología computacional de la red alianza suma de Manizales, de reciente creación.

En BIOS hemos identificado la necesidad de estimular la oferta y demanda del sector biotecnológico, fortaleciendo la cultura Bio apoyada con TIC, mediante la difusión, la formación de alto nivel y la transferencia tecnológica. El Centro ofrece permanentemente cursos de corta duración en de análisis e interpretación de datos biológicos, visualización de anotación de genomas, estadística en el análisis de datos genómicos y diversas áreas de la biotecnología y bionegocios. BIOS cuenta además con un portafolio de 7 diplomados relacionados con las líneas de investigación del centro y participa como aliado en la maestría en bioinformática y biología computacional de la red alianza suma de Manizales, de reciente creación.

FUTUROS RETOS

El futuro de BIOS es muy promisorio. Contamos con la computadora para análisis de datos científicos más potente del país, con solo el 10 por ciento de la capacidad instalada, es decir, que en muy poco tiempo, con nuestro 100 por ciento, contaremos con la herramienta más robusta para este tipo de análisis de todo Sur América. Los proyectos emblemáticos que está desarrollando el centro le permitirá convertirse en líder nacional en la creación de los estándares en la toma de datos de las ciencias "-ómicas", los análisis de los mismos, el almacenamiento

de esta gran cantidad de información y deber ser el foco central de las plataformas de visualización y acceso de los datos para todos los grupos de investigación involucrados.

Los ambiciosos proyectos de BIOS hacen vislumbrar para el país un crecimiento económico de su agroindustria y sector salud basado en innovación científica que permita la creación de nuevos mercados y la optimización de los existentes dentro de un contexto sostenible.

REFERENCIAS

[1] A. Von HUMBOLDT, "La biodiversidad y los servicios ecosistémicos." [Online]. Available: <http://humboldt.org.co/biodiversidad>. [Accessed: 05-Jun-2014].

[2] "Consejo Nacional de Política Económica y Social - CONPES 3697 Política para el Desarrollo Comercial de la Biotecnología a partir del uso Sostenible de la Biodiversidad," 2011.

[3] D. N. de P.- DNP, "Colombia adopta política para el desarrollo comercial de la biotecnología.," 2011.

[4] A. P. Arango, G. De Análisis, and J. D. Vargas, "Plan Estratégico Nacional de Mercados Verdes Ministerio del Medio Ambiente," no. 8, 2002.

[5] O. for E. C. and Development, "The Bioeconomy to 2030 Main Findings and Policy Conclusions," 2009.

[6] L. Walker, "Building on Global Drivers of Innovation," *Ind. Biotechnol.*, vol. 8, no. 4, pp. 163 –163, 2012.

[7] C. B. y C. R. MELGAREJO, L. M., J. SÁNCHEZ, A. CHAPARRO, F. NEWMARK, M. SANTOS-ACEVEDO, "Aproximación al estado actual de la bioprospección en Colombia." *Cargrapics*, p. 334, 2002.

[8] S. A. LAIRD, *Introduction: equitable partnerships in practice.*, Sarah A. L. London: Biodiversity and Traditional Knowledge. *Equitable Partnerships*, 2002, pp. xxii–xxxvi.

[9] O. D. Torres and L. Velho, "La bioprospección como un mecanismo de cooperación internacional para fortalecimiento de capacidades en ciencia y tecnología en Colombia," *Ciência da Informação*, vol. 38, no. 3, pp. 96–110, Dec. 2009.

[10] J. M. Feinsilver, "Prospección de la biodiversidad: potencialidades para los países en desarrollo," *CEPAL*, vol. 60, pp. 111–128, 1996.

[11] L. U. Z. M. Melgarejo, "Bioprospection : National Plan and Up to Date Situation in Colombia," *Acta biológica Colomb.*, vol. 8, no. 2, 2003.



PROYECTOS EMBLEMÁTICOS

PARA FORTALECER LA BIOINFORMÁTICA Y LA BIOLOGÍA COMPUTACIONAL EN COLOMBIA

*Ángela María Rojas Sánchez, Marco Aurelio Cristancho,
Dago Hernando Bedoya Ortiz*
Autores

La bioinformática es una disciplina científica que logró beneficiar los procesos industriales, médicos y agronómicos del mundo al unir la genética con la informática y computación en una nueva revolución en la que los datos y la vida construyen nuevas formas de leer e interpretar al mundo. Este compendio busca presentar las iniciativas colombianas para el futuro en este campo, luego de realizar el Encuentro para la Identificación de Proyectos Emblemáticos y Conformación de una Comunidad Bioinformática, en la que expertos nacionales e internacionales se cuestionaron sobre cuáles deben ser los esfuerzos conjuntos para convertir la bioinformática y todas las disciplinas que están a su alrededor, en un motor socio-económico para Colombia.

Durante este encuentro los expertos analizaron las fortalezas y debilidades de esta disciplina en Colombia, las posibilidades de formación de alto nivel y de encaminar los esfuerzos conjuntos a iniciativas en las que todos puedan participar, se definieron 12 propuestas de proyectos nacionales, los cuales fueron calificados por su viabilidad y capacidad de generar conexión interinstitucional. El Genoma Colombiano y la Expedición Botánica in silico fueron las dos iniciativas escogidas, en las que el país, como comunidad científica, debe abordar y desarrollar de manera conjunta. La ciencia colaborativa generará grandes frutos para la prosperidad de las próximas generaciones.

MIEMBROS BIOS



Teléfono: (+057) (6) 8755656 | 8930360 | 8753799
Email: contacto@bios.co
www.bios.co
Dirección: Ecoparque los Yarumos, Edificio BIOS
Manizales, Caldas, Colombia



CENTRO DE BIOINFORMÁTICA
Y BIOLOGÍA COMPUTACIONAL