# LA REVOLUCIÓN DEL BIG DATA

¿Cómo podríamos usar estos datos para pronosticar y mejorar nuestras investigaciones?

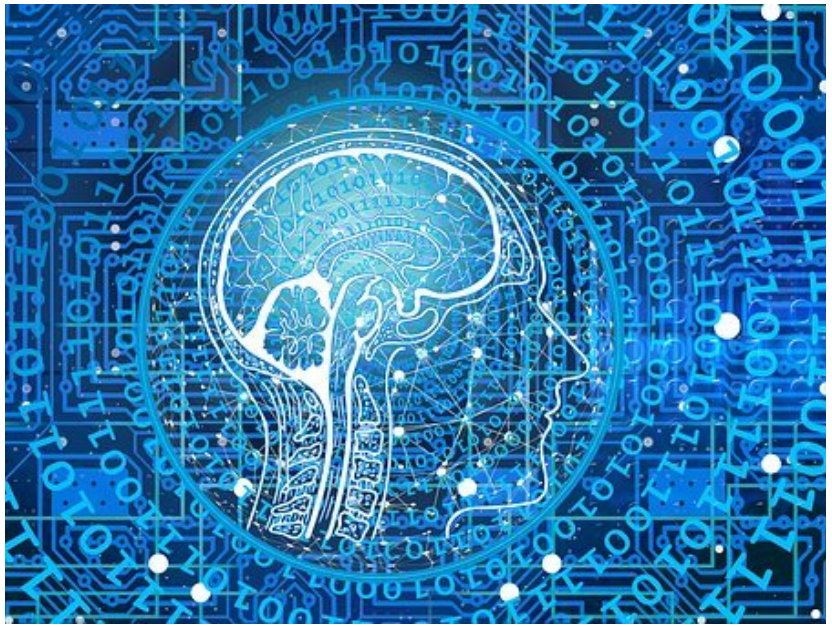
UNIVERSIDAD DE ANTIOQUIA

Héctor Alejandro Rodríguez Cabal PhD
Docente FCEN
Coordinador del Grupo Agrobiotecnología

# Contenido

UNIVERSIDAD DE ANTIOQUIA

- Concepto de Big Data.
- Diferencias entre Big Data y Data Science.
- Big Data en el mundo.
- Big Data en Colombia.
- Big Data para la ciencia.
- Perspectivas.



# ¿Qué es Big Data?

UNIVERSIDAD DE ANTIOQUIA
1803

https://www.powerdata.es/big-data

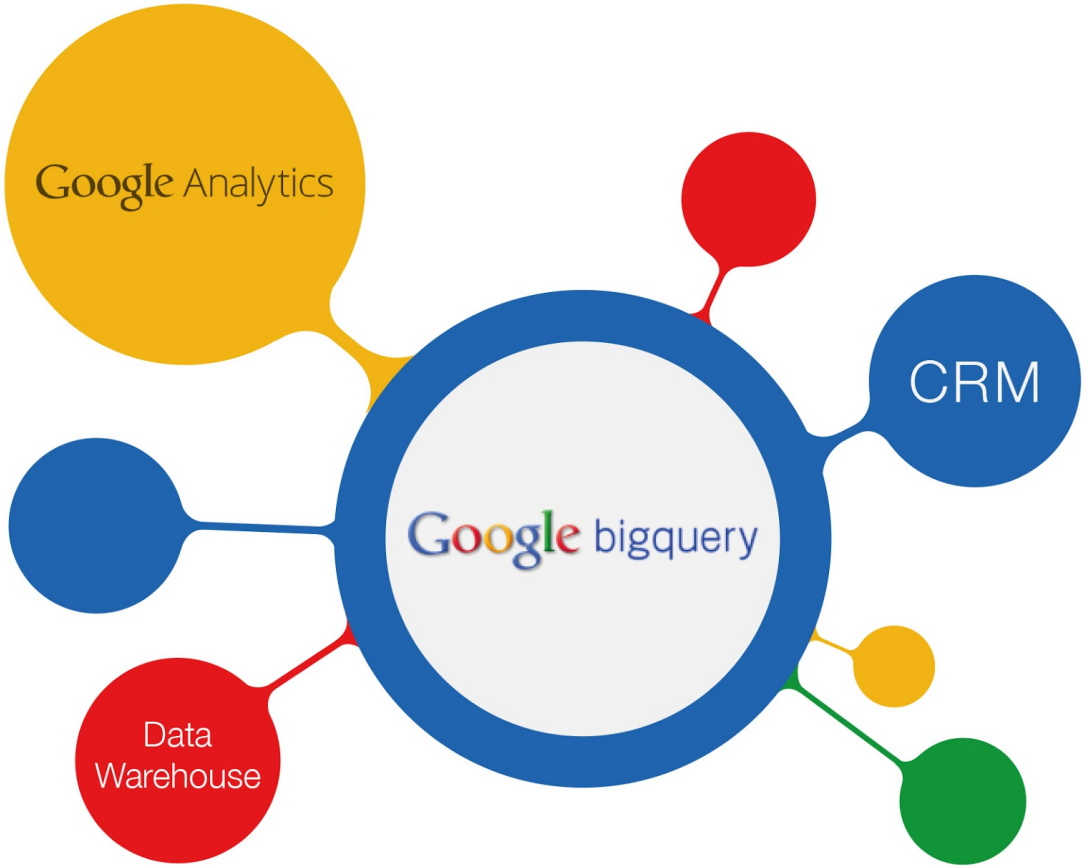
- Web logs.
- La identificación por radiofrecuencia (RFID) Los sensores incorporados en dispositivos.
- La maquinaria.
- Los vehículos.
- Las búsquedas en Internet.
- Las redes sociales como Facebook.
- Computadoras portátiles.
- Teléfonos inteligentes y otros teléfonos móviles.
- Dispositivos GPS.
- Registros de centros de llamadas.

# ¿Por qué el Big Data es tan importante?

- El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades.

  - **Reducción de coste**.
  - **Más rápido, mejor toma de decisiones.**
  - **Nuevos productos y servicios**.

# El as en la manga de big data Google




Si Google encuentra que de repente hay una cantidad creciente de usuarios que empieza a mirar determinado tipo de contenido concreto, es posible que esté detectando una nueva tendencia antes de que se convierta en popular.

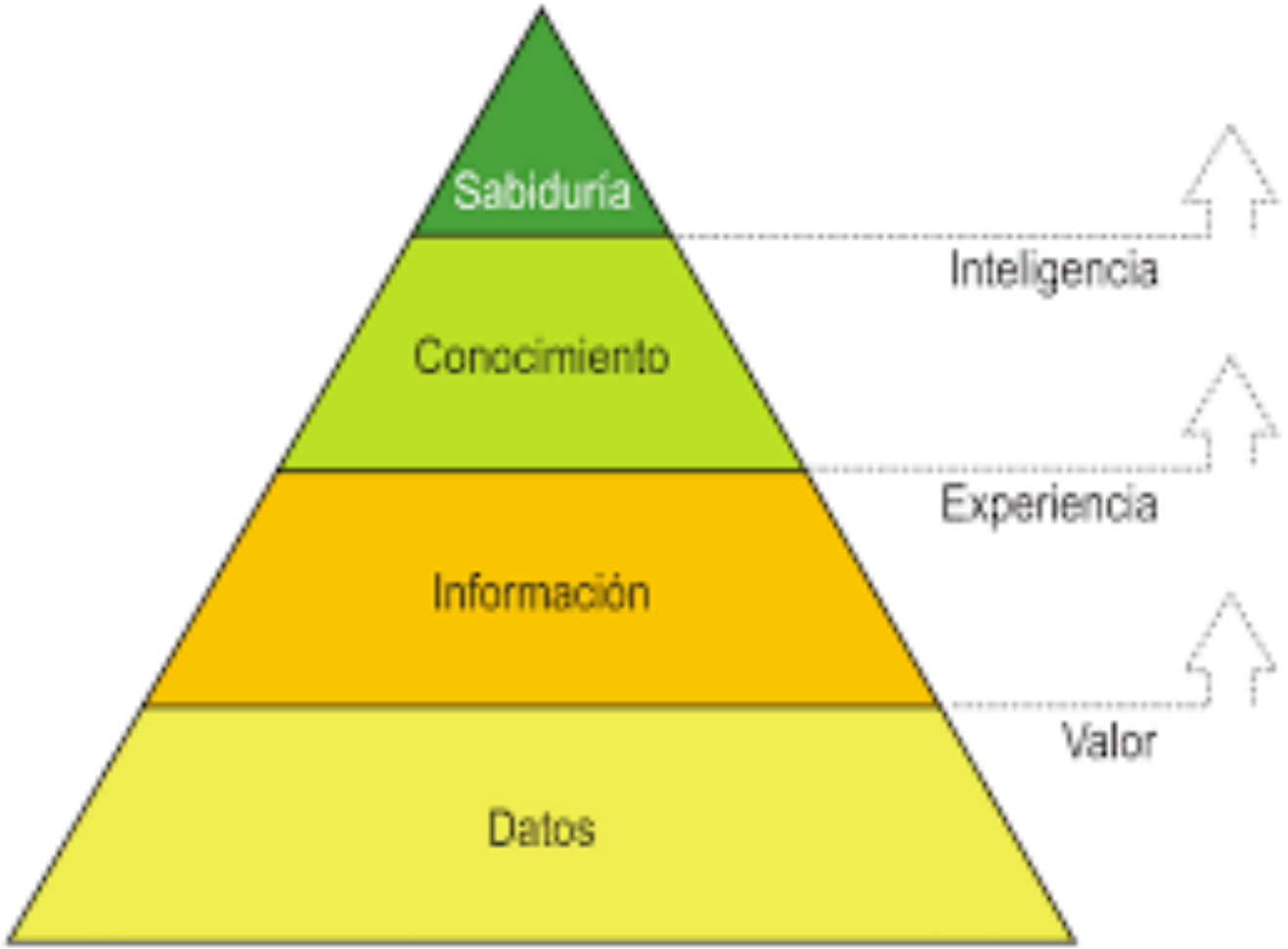
"Con el big data somos capaces de decirle a un restaurante dónde es mejor abrir para tener éxito", Susana Voces de Deliveroo

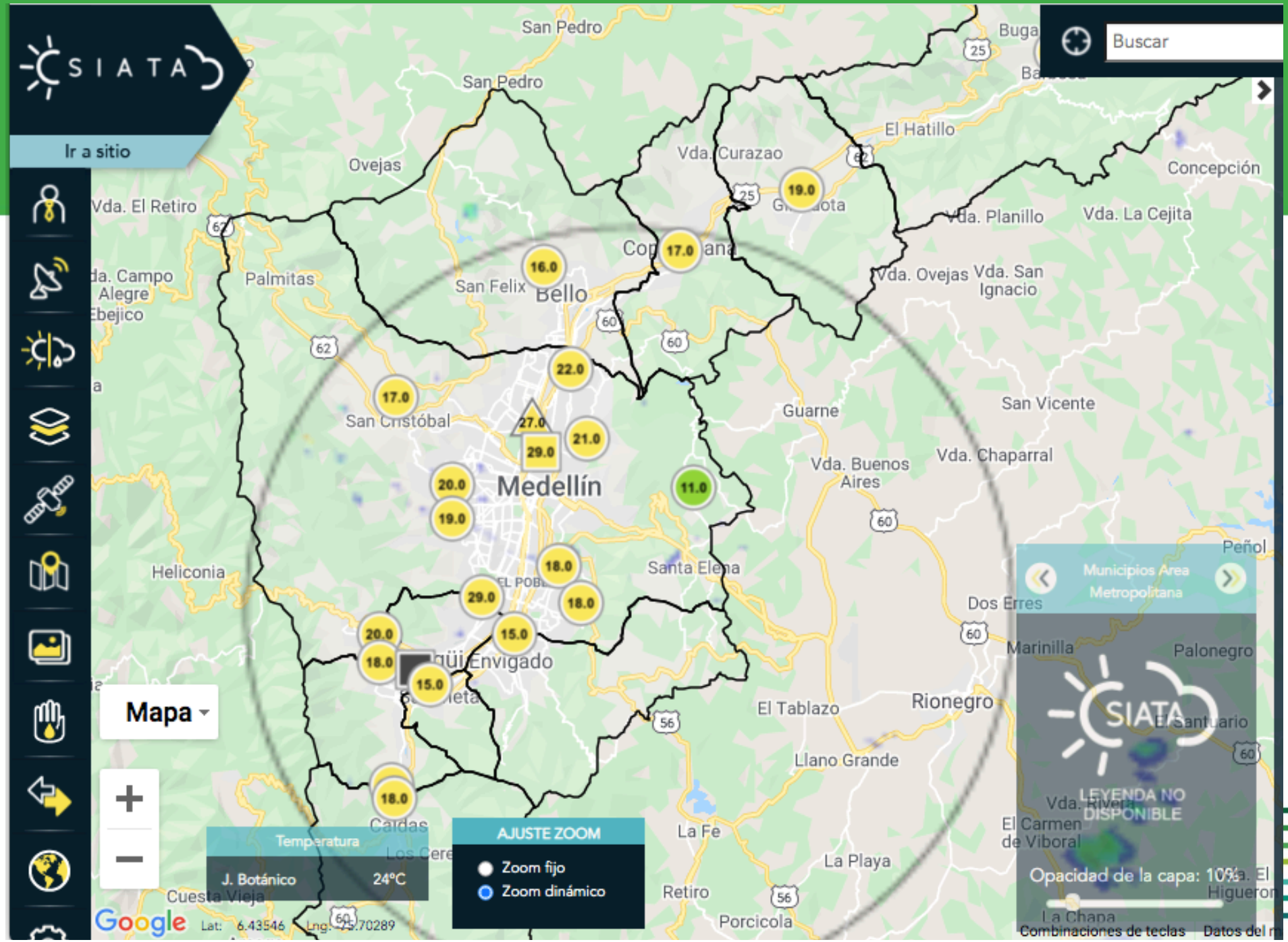
# Diferencias entre Big Data y Data Science.

Big Data se distinguen por variedad, volumen y velocidad

El Data Science proporciona métodos o técnicas para analizarlos.

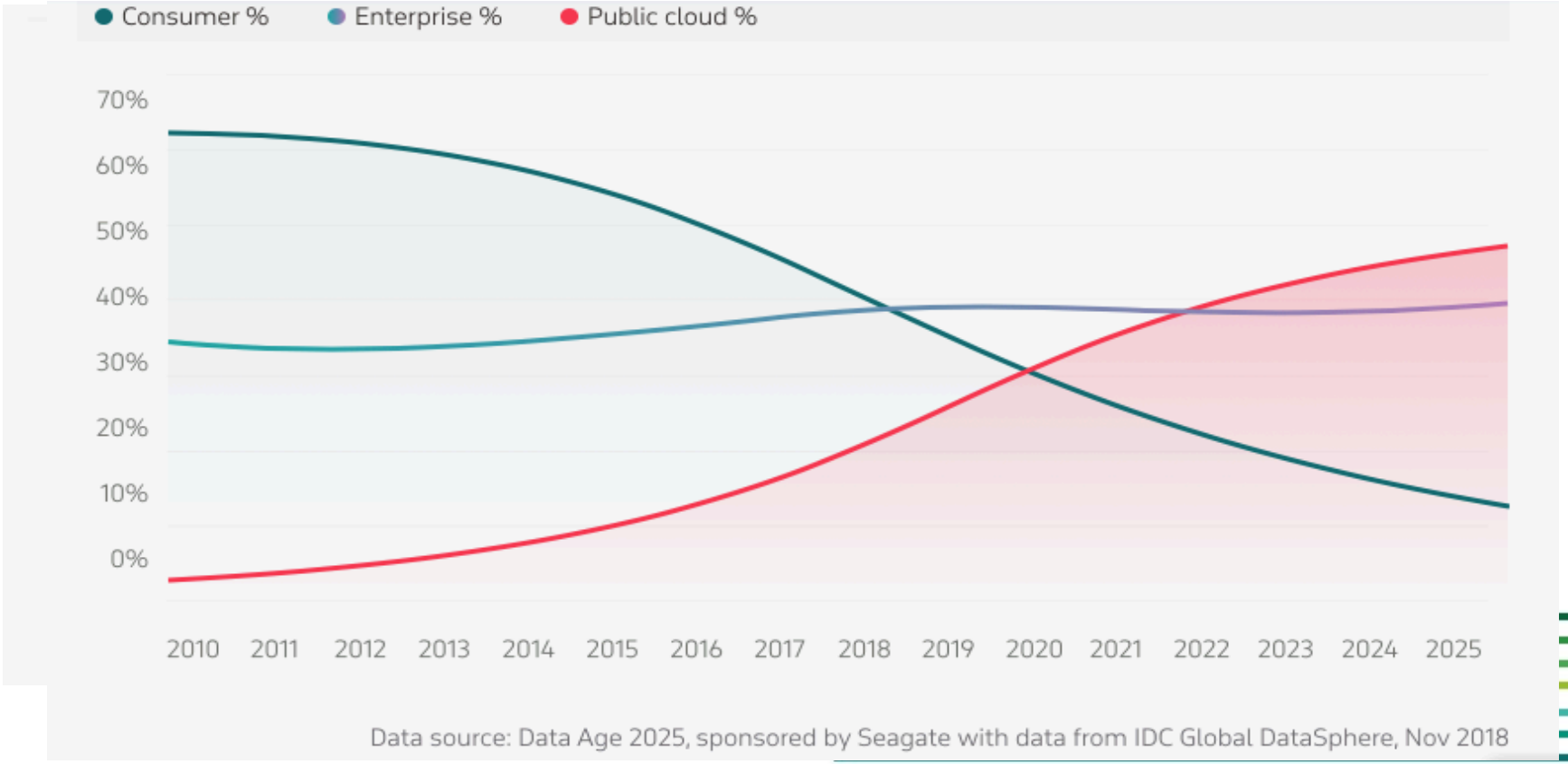
# Pirámide informacional




SIATA
Ir a sitio
Buscar
Mapa
Medellín
Bello
San Cristóbal
Envigado
Caldas
Rionegro
Guarne
Santa Elena
Heliconia
Palmitas
San Pedro
Ovejas
El Hatillo
Vda. Curazao
Concepción
Vda. La Cejita
Vda. Planillo
Vda. Ovejas
Vda. San Ignacio
San Vicente
Vda. Chaparral
Vda. Buenos Aires
Peñol
Dos Erres
Marinilla
Palonegro
El Santuario
El Tablazo
Llano Grande
La Fe
La Playa
Retiro
Porcicola
El Carmen de Viboral
La Chapa
Cuesta Vieja
Vda. El Retiro
Vda. Campo Alegre
Municipios Area Metropolitana
LEYENDA NO DISPONIBLE
Opacidad de la capa: 100%
Temperatura
J. Botánico 24°C
AJUSTE ZOOM
Zoom fijo
Zoom dinámico
Google
Lat: 6.43546
Combinaciones de teclas
Datos del m

# Big data en el mundo

1. Los volúmenes de datos seguirán aumentando y migrando a la nube



# Big data en el mundo

2. El aprendizaje automático (machine learning) seguirá cambiando el panorama.

> **Machine learning is becoming more sophisticated with every passing year. We are yet to see its full potential—beyond self-driving cars, fraud detection devices, or retail trends analyses.**

**Wei Li**

Vice President and General Manager at Intel

> **What fascinates me is combining big data with machine learning and especially natural language processing, where computers do the analysis by themselves to find things like new disease patterns.**

**Bernard Marr**

Author, Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance

# Big data en el mundo

## 3. La privacidad seguirá siendo un tema candente

- **Brecha de habilidades de seguridad**, causada por la falta de oportunidades de educación y capacitación. Esta brecha crece constantemente y alcanzó los 3,5 millones de puestos de seguridad cibernética vacantes para 2021, según Cybercrime Magazine.
- **Evolución de los ciberataques.** Las amenazas utilizadas por los piratas informáticos están evolucionando y se vuelven más complejas cada día.
- **Adhesión irregular a las normas de seguridad.** Aunque los gobiernos están tomando medidas para estandarizar las regulaciones de protección de datos, siendo GDPR el ejemplo, la mayoría de las organizaciones aún ignoran los estándares de seguridad de datos.

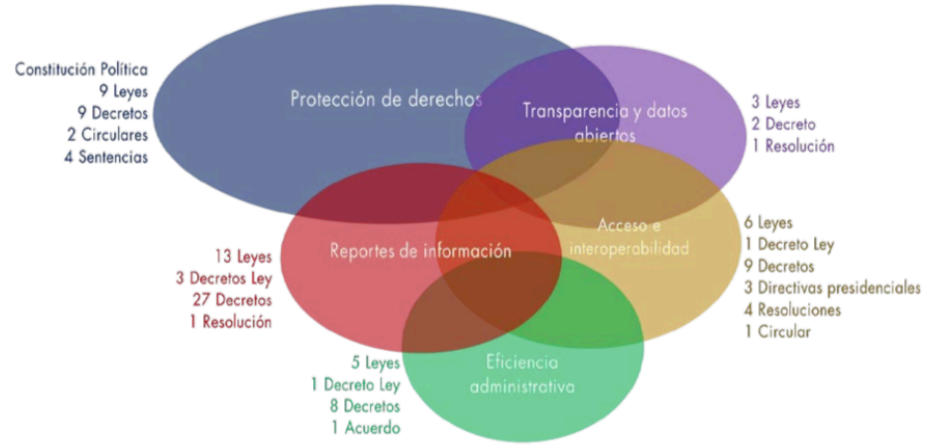
# Big data en el mundo

4. Los datos rápidos y procesables pasarán a primer plano

# Big Data en Colombia.

El Gobierno de Colombia ha realizado importantes esfuerzos para consolidar un marco de política que promueva el aprovechamiento de datos para la generación de valor social y económico en el país. La Política Nacional de Explotación de Datos y Big data - CONPES 3920 (2018) y la Política Nacional de Transformación Digital e Inteligencia Artificial CONPES 3975 (2019).

**Figura 1. Marco jurídico aplicable a los datos**




Fuente: DNP (2018).

# Big Data en Colombia.

## MinTIC expide el Plan Nacional de Infraestructura de Datos, que impulsará la transformación digital del Estado

Nacional

# Enfoque de datos como infraestructura

- Permite abordar la infraestructura de datos, desde un enfoque de bien público, transversal a diversas actividades del Estado, necesario para crear otros bienes y servicios en la economía.

# Propósitos del Plan Nacional de Infraestructura de Datos

- Aumentar la reutilización de los datos que integran la infraestructura de datos
- Consolidar un sector público basado en datos
- Consolidar espacios de intercambio de datos que impulsen la innovación en el país
- Promover el desarrollo e integración de tecnologías emergentes
- Posicionar al modelo de gobernanza de los datos para la consolidación de una economía digital guiada por datos
- Posicionar al país como un referente en el uso de los datos para el desarrollo de la economía digital
- Construir un entorno de confianza pública para el aprovechamiento y protección de los datos

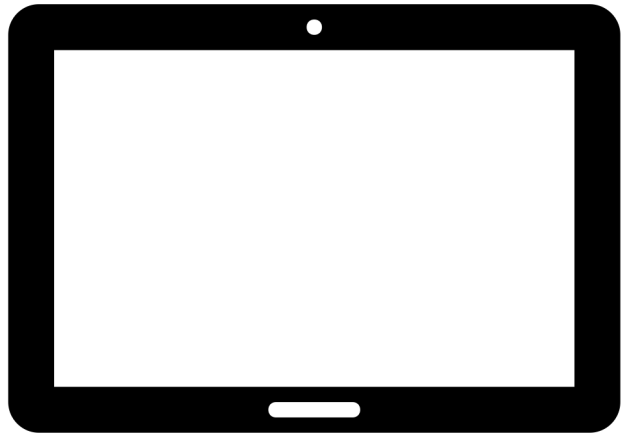
# Big Data en Colombia.

GOV.CO

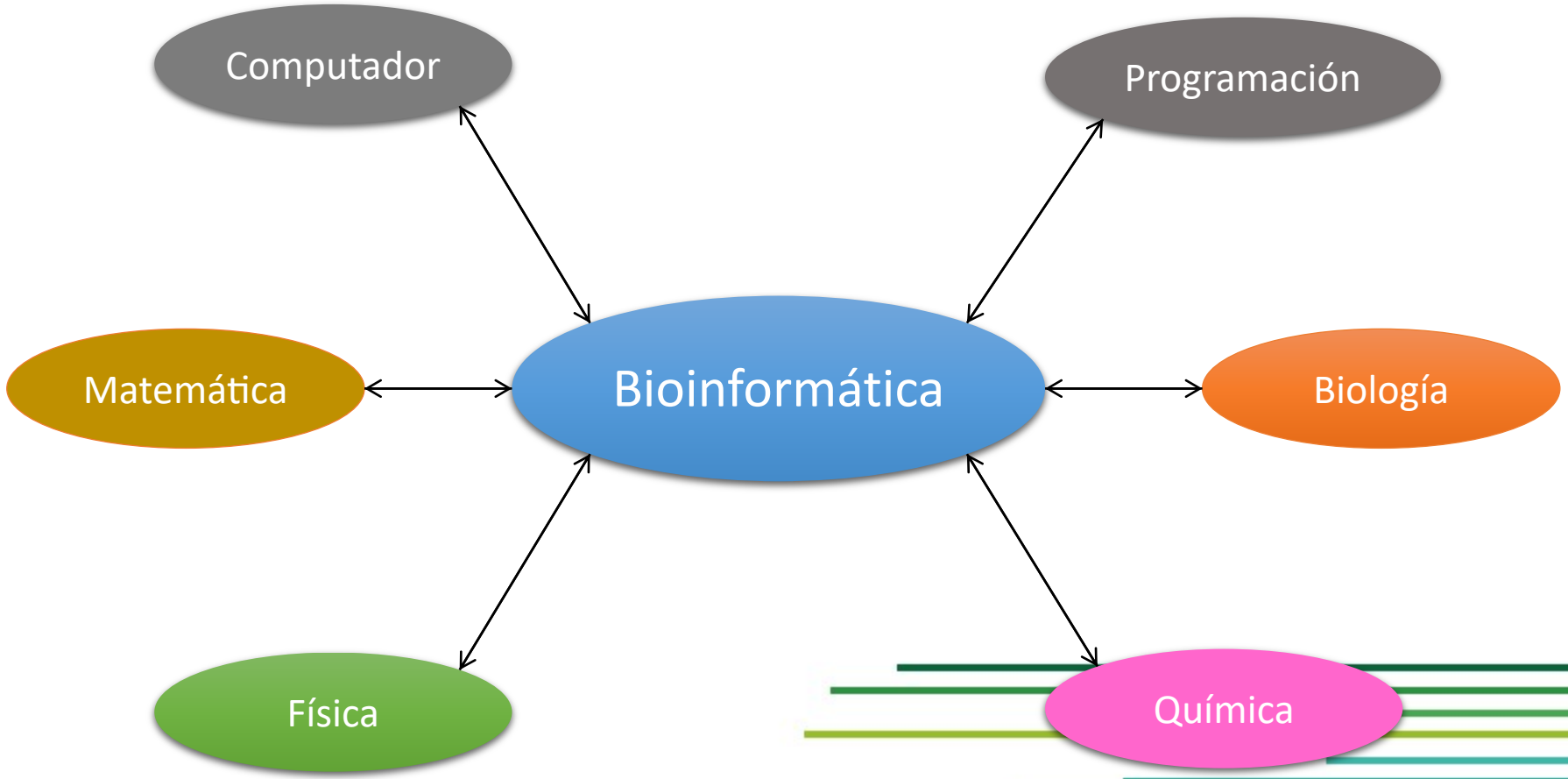
**Misión de Expertos de Inteligencia Artificial**

La Misión de Expertos en IA de Colombia es uno de los primeros grupos de expertos de IA en la región, y uno de los primeros en enfocarse en la generación de medidas para el desarrollo de políticas de educación y empleo para la cuarta revolución industrial.

# ¿Podria Big Data impactar la educación?



# Big data para la Ciencia

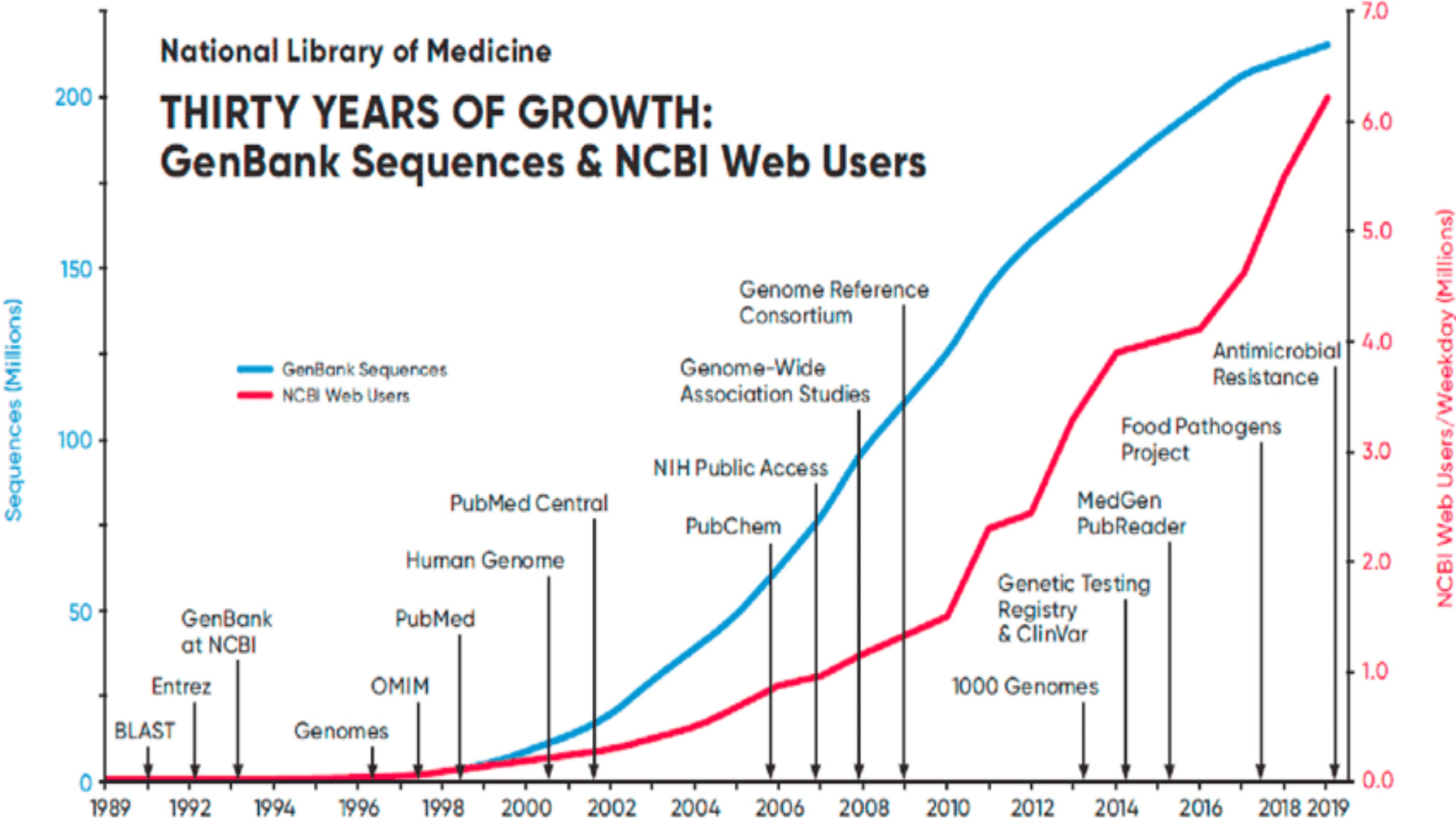


# Big data para la Ciencia

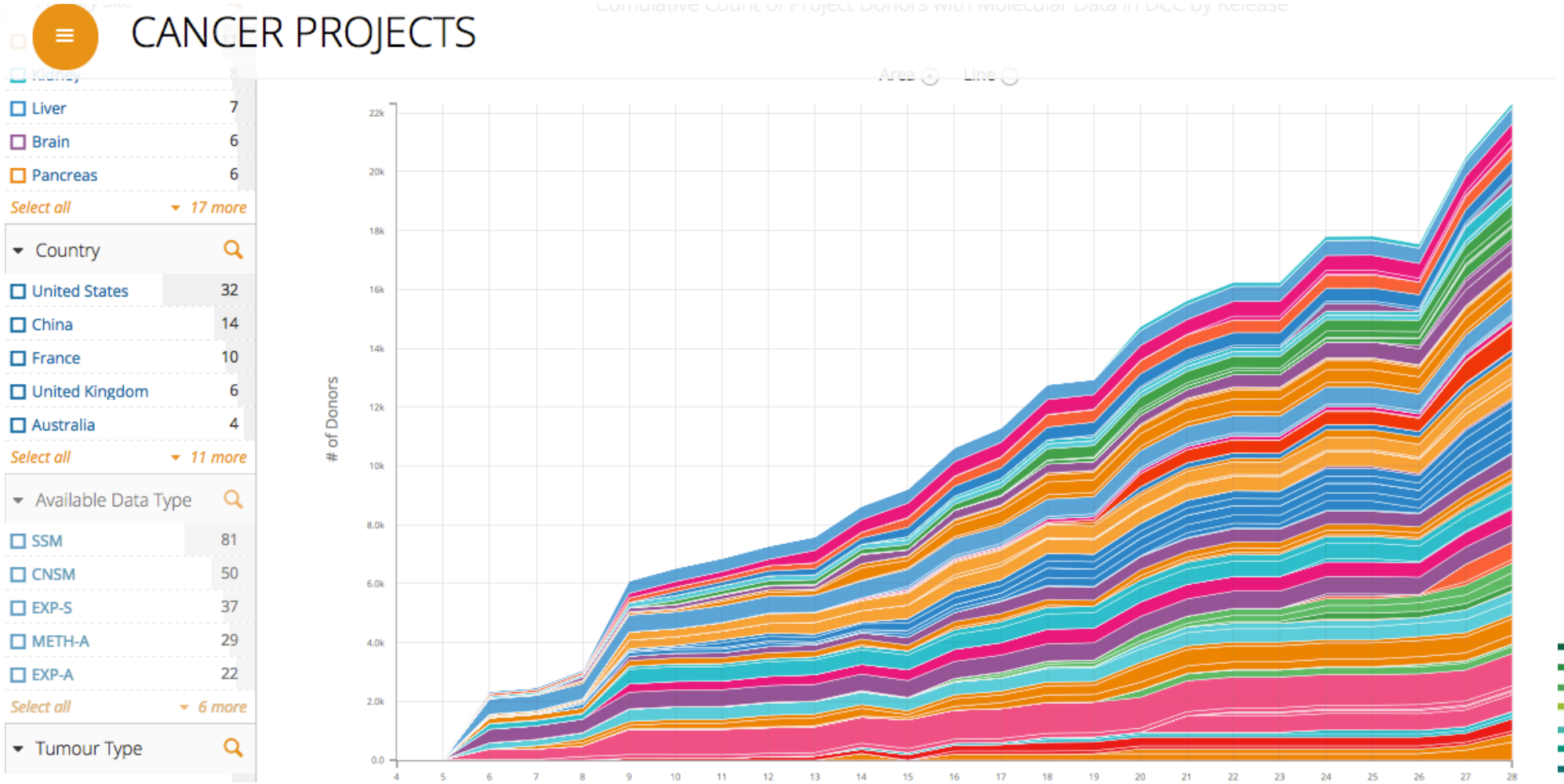
## En proyectos a gran escala

- Se busca disminuir al máximo la intervención humana.
- Sistematización del proceso.
- Diseño de pipelines y nuevos scripts

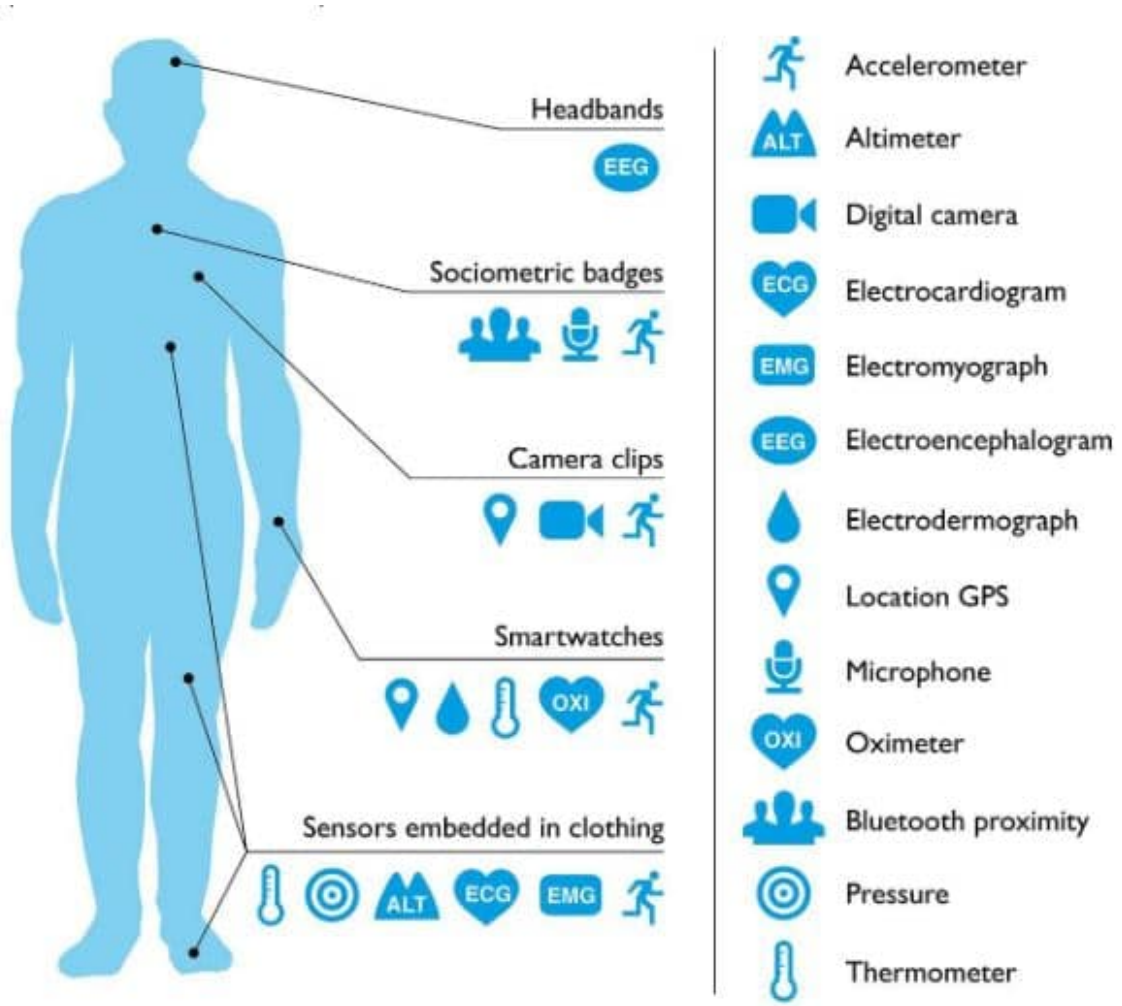
# Big data para la Ciencia


National Library of Medicine
THIRTY YEARS OF GROWTH:
GenBank Sequences & NCBI Web Users
GenBank Sequences
NCBI Web Users
Sequences (Millions)
NCBI Web Users/Weekday (Millions)
0
50
100
150
200
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
1989
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2019
BLAST
Entrez
GenBank at NCBI
Genomes
OMIM
PubMed
Human Genome
PubMed Central
PubChem
NIH Public Access
Genome-Wide Association Studies
Genome Reference Consortium
1000 Genomes
Genetic Testing Registry & ClinVar
MedGen PubReader
Food Pathogens Project
Antimicrobial Resistance

# Big data para la Ciencia



# Big data para la Ciencia



# Big data para la Ciencia

ADVERTISEMENT FEATURE Advertiser retains sole responsibility for the content of this article

## Startups and big data set to fuel genomic medicine in Korea

Genomic medicine is taking off in South Korea in a big way, thanks in large part to innovative startups and systems for collecting and analyzing big data.

---

Produced by

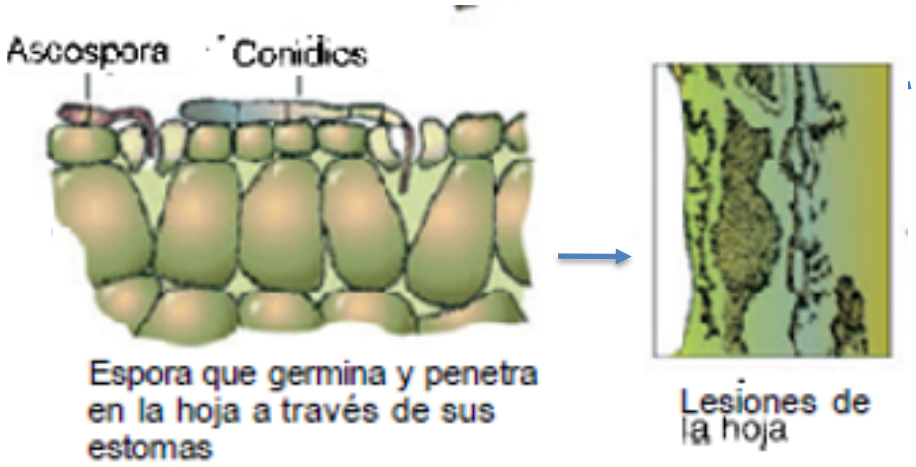
nature research custom media

SEOUL NATIONAL UNIVERSITY

Variedad Susceptible Williams

Variedad resistente Calcutta 4

# Apoplastic fungus



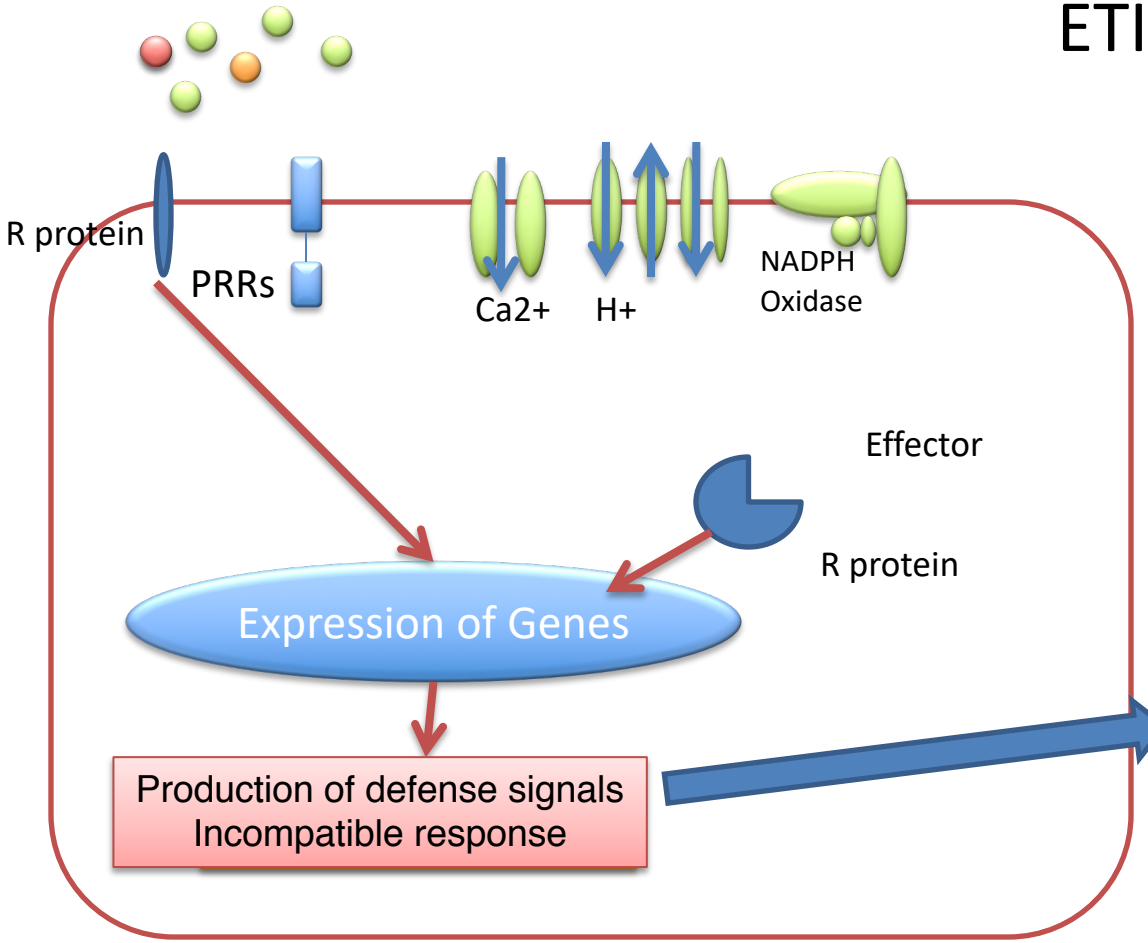

Biotrophic → Necrotrophic

10 to 15 days

Long biotrophic phase

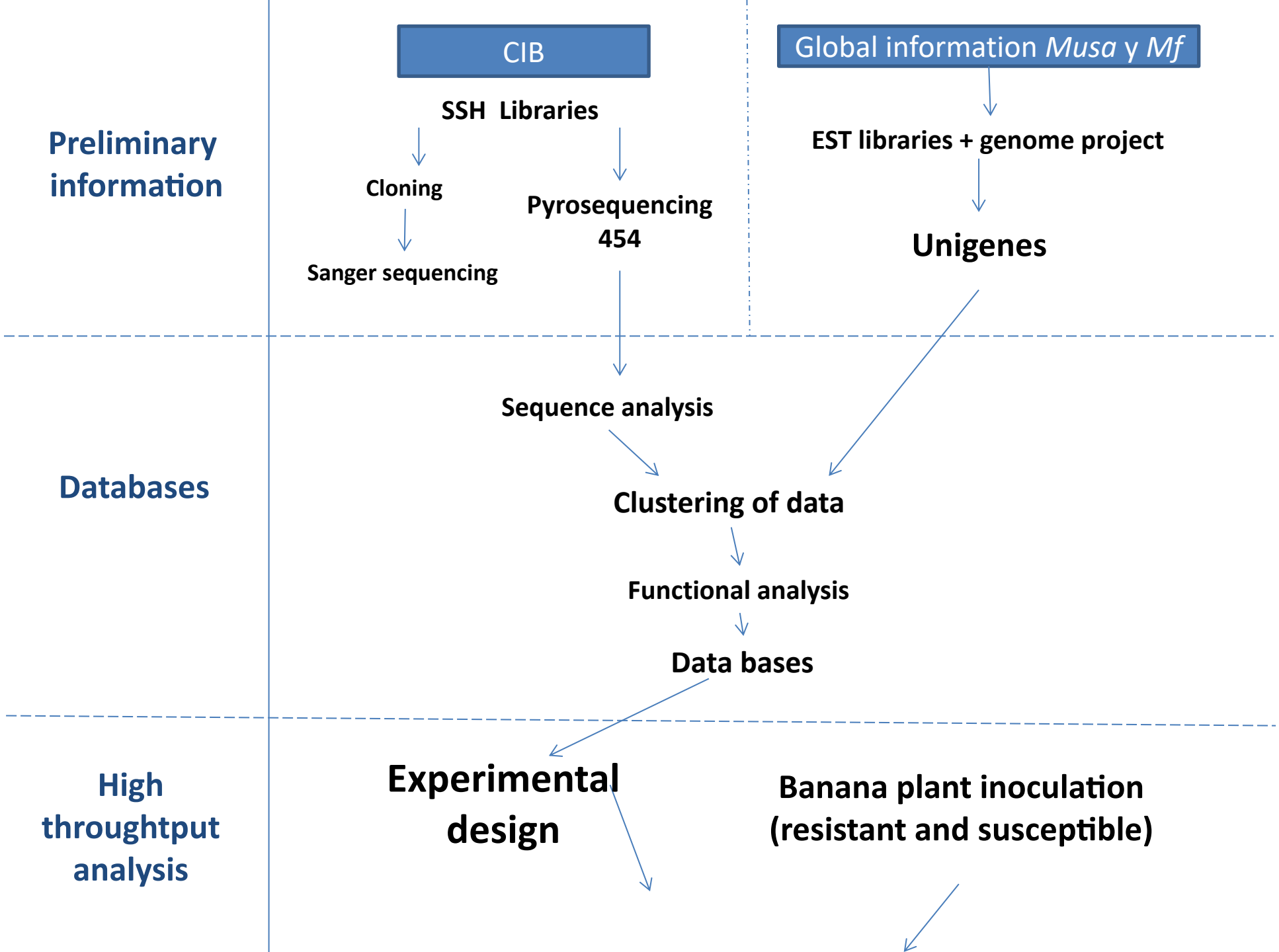
# How Plants Defend Themselves?

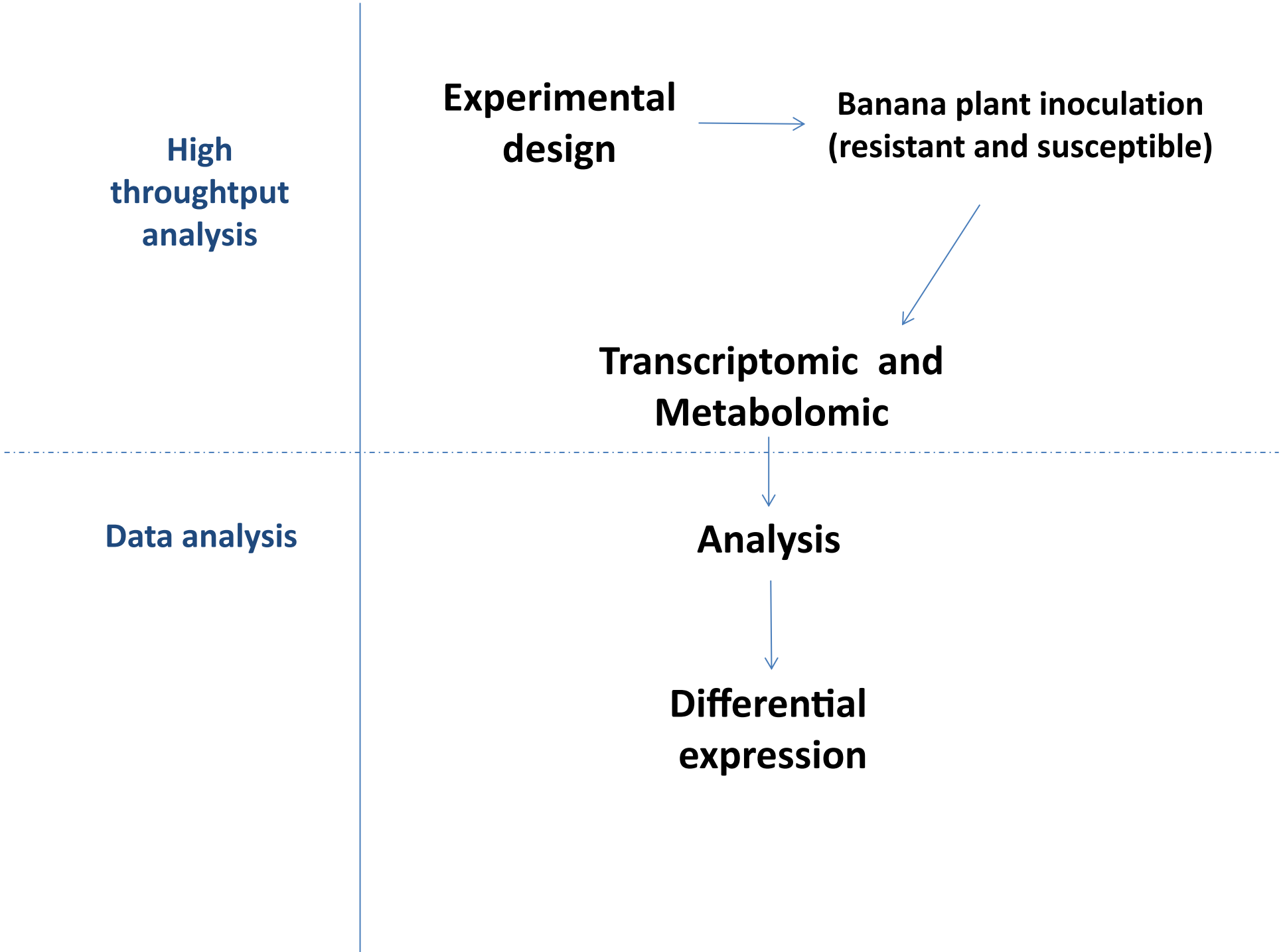
PTI




ETI

- -Oxidative pathways
- -Production of ROS
- -Transcription factors
- -Hormone signals
- -Phenoloxidase, phenylphenalenones
- -Cell death
- -Systemic adquired resistance (SAR)


Preliminary information
Databases
High throughtput analysis
CIB
Global information *Musa* y *Mf*
SSH Libraries
EST libraries + genome project
Cloning
Pyrosequencing 454
Sanger sequencing
Unigenes
Sequence analysis
Clustering of data
Functional analysis
Data bases
Experimental design
Banana plant inoculation (resistant and susceptible)


High throughtput analysis
Experimental design
Banana plant inoculation (resistant and susceptible)
Transcriptomic and Metabolomic
Data analysis
Analysis
Differential expression

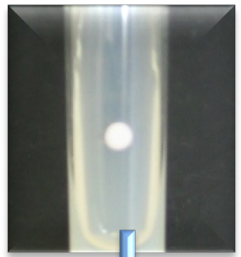
| 12 hpi | 18 hpi | 24 hpi | 48 hpi | 72 hpi | 144 hpi | 9 dpi | 15 dpi | 30 dpi |
|---|---|---|---|---|---|---|---|---|
| control1 | control1 | control1 | control1 | control1 | control1 | control1 | control1 | control1 |
| control2 | | control2 | | control2 | control2 | | control2 | |
| Control 3 | | Control 3 | | Control 3 | Control 3 | | Control 3 | |
| Replica 1 | Replica 1 | Replica 1 | Replica 1 | Replica 1 | Replica 1 | Replica 1 | Replica 1 | Replica 1 |
| Replica 2 | | Replica 2 | | Replica 2 | Replica 2 | | Replica 2 | |
| Replica 3 | | Replica 3 | | Replica 3 | Replica 3 | | Replica 3 | |
| Contact | Contact | Penetration | Advance | Advance | HR (resistant) | Stage 0-1 susptible | Stage 2-3 susceptible | Stage 4-5 susceptible |

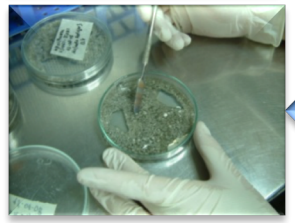
# Inoculation

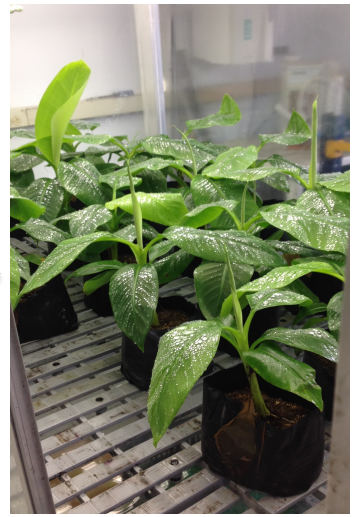
Williams (Susceptible)

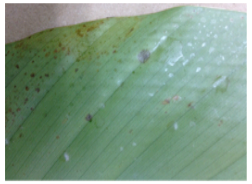
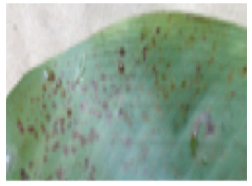
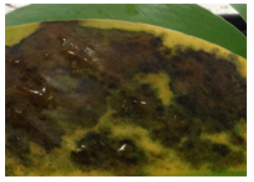


Calcutta (Resistente)



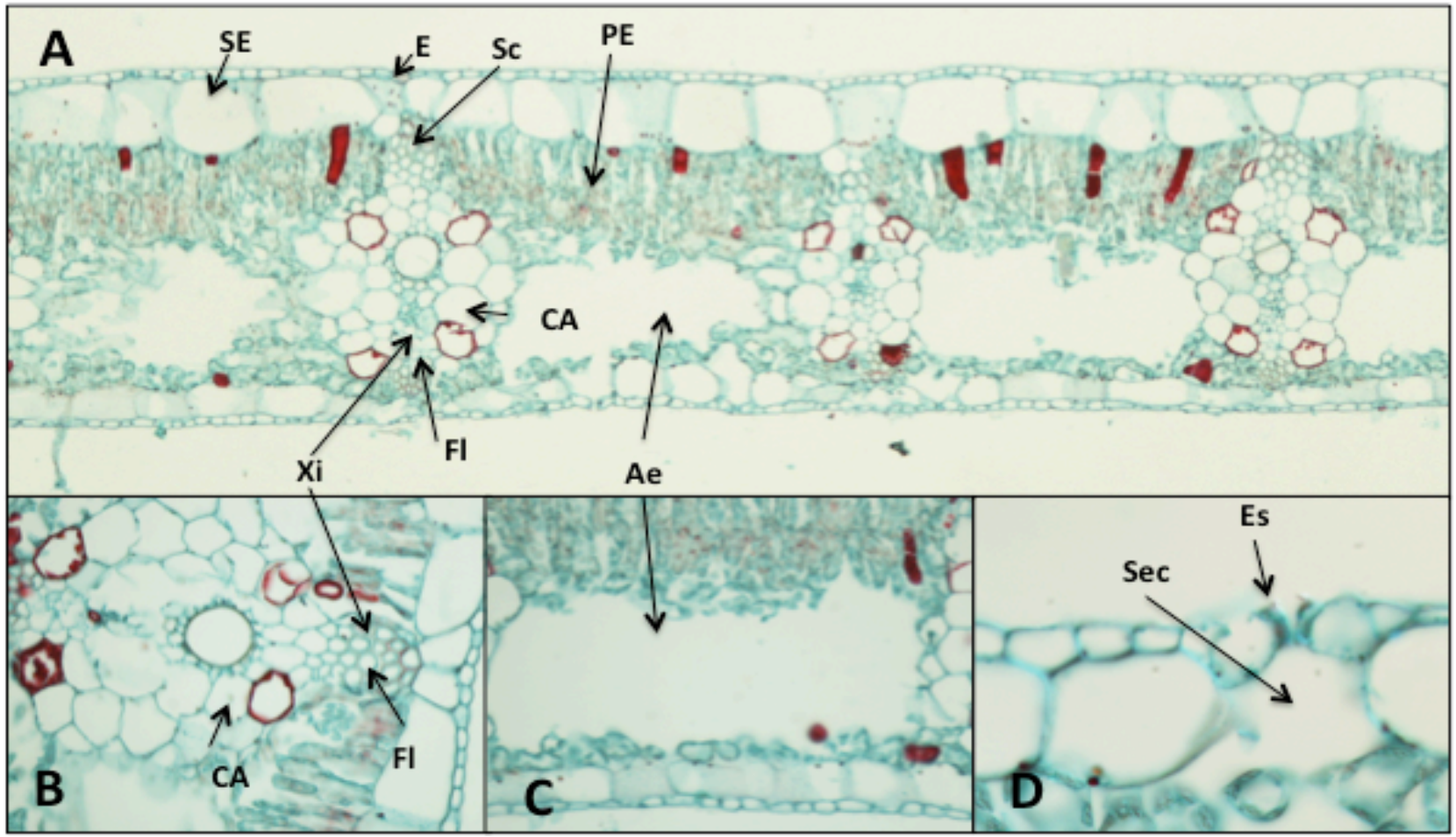




| EST | E0 | E1/2 | E2 | E3 | E3 | E3 | E3/4 |
|---|---|---|---|---|---|---|---|
| T | 0 - 72 hpi | 144 hpi (6 dpi) | 216 hpi (9 dpi) | 360 hpi (15 dpi) | 552 hpi (23 dpi) | 720 hpi (30 dpi) | 960 hpi (40 dpi) |
| CAL |  |  |  |  |  |  |  |
| EST | E0 | E1 | E2 | E3 | E4 | E5 | E6 |
| T | 0 - 216 hpi | 240 hpi (10 dpi) | 312 hpi (13 dpi) | 360 hpi (15 dpi) | 552 hpi (23 dpi) | 720 hpi (30 dpi) | 960 hpi (40 dpi) |
| Will |  |  |  |  |  |  |  |

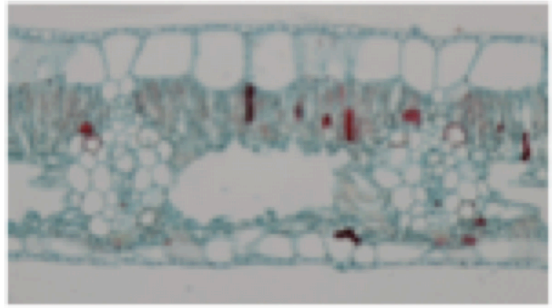
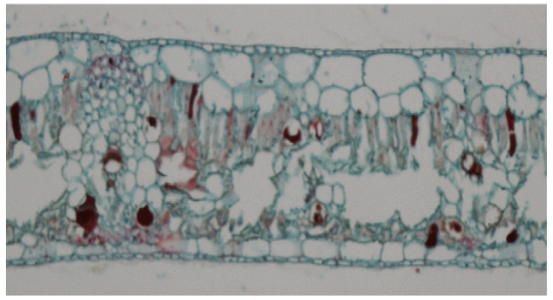
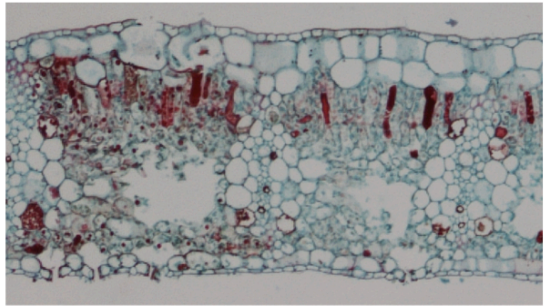
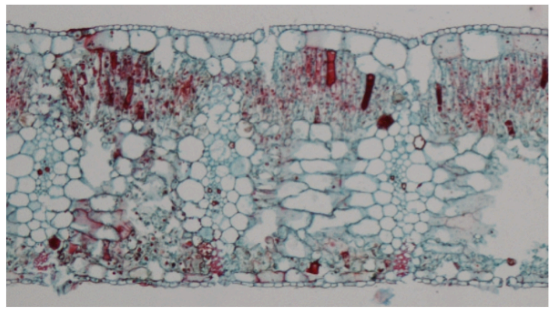
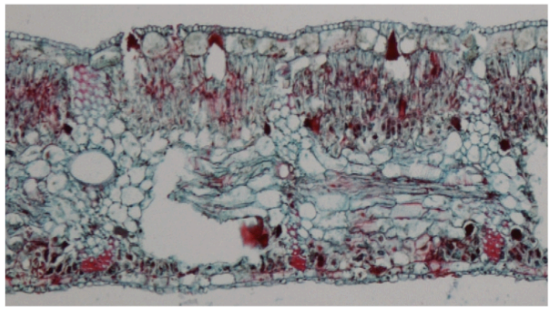
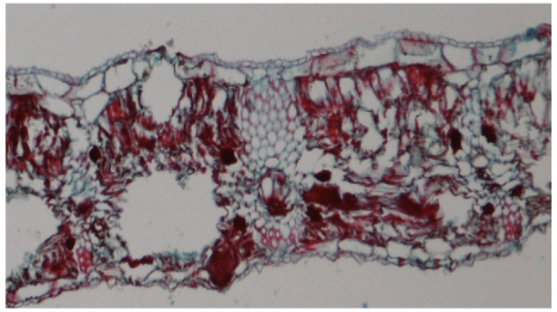
Est= Estadío; T= tiempo

# Caracterización morfoanatómica de una hoja sana de Williams



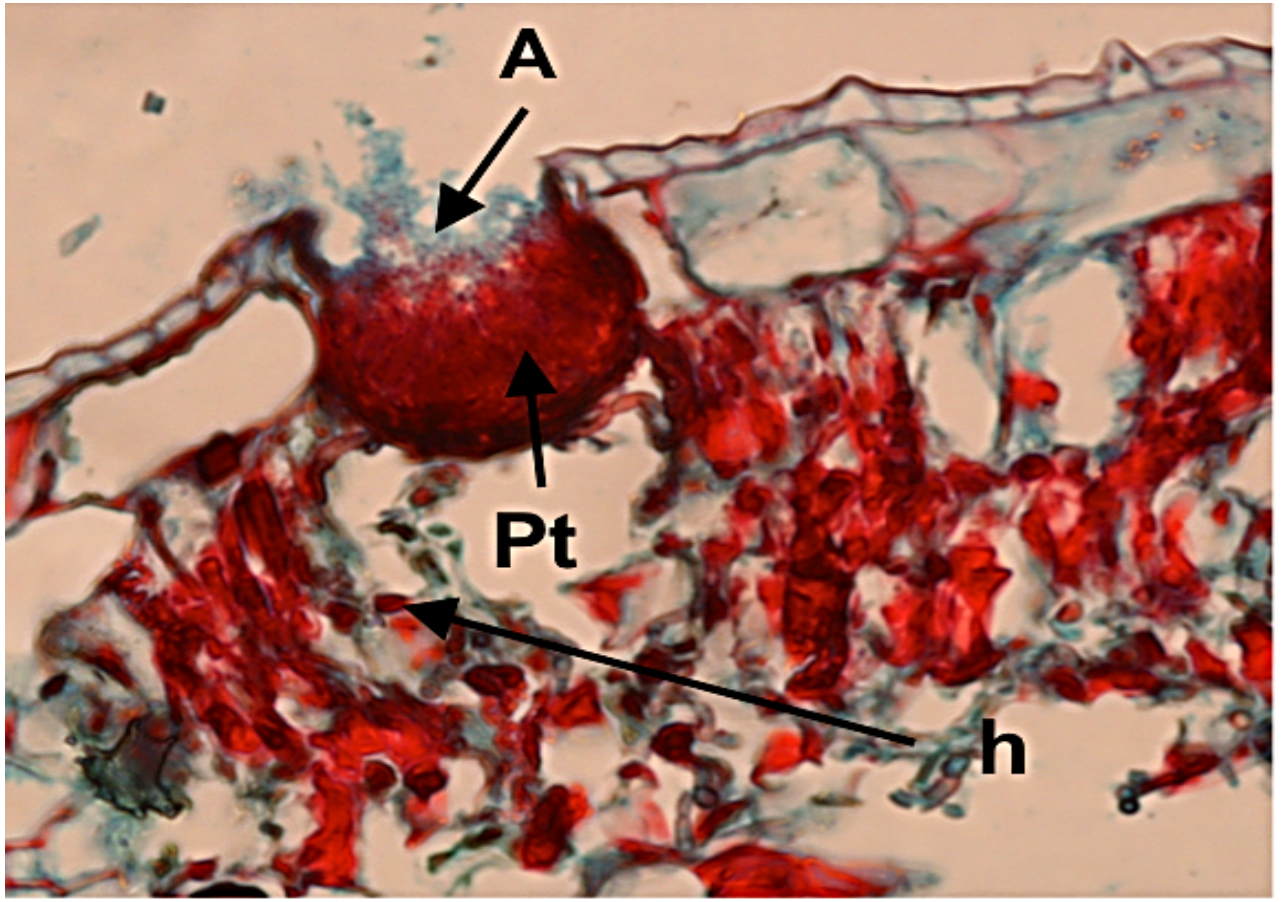

SE= capa Sub-Epidermal; PE= células del Parénquima de Empalizada; E= Epidermis; Sc= células del Esclerénquima; CA=Células Acompañantes; Sec= cámara Sub-estomatal; Ae= Espacios aeríferos; Fl= Floema; Es= Estoma; Xi= Vasos del Xilema.

# Caracterización morfoanatómica de hojas de la variedad Williams después de la inoculación con el hongo.

| EST | E1 | E2 | E3 |
|---|---|---|---|
| T | 240 hpi (10 dpi) | 312 hpi (13 dpi) | 360 hpi (15 dpi) |
| Will |  |  |  |
| **EST** | **E4** | **E5** | **E6** |
| **T** | **552 hpi (23 dpi)** | **720 hpi (30 dpi)** | **960 hpi (40 dpi)** |
| Will |  |  |  |

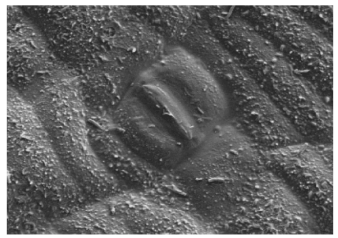
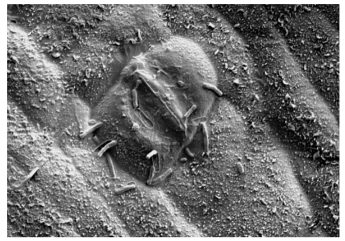
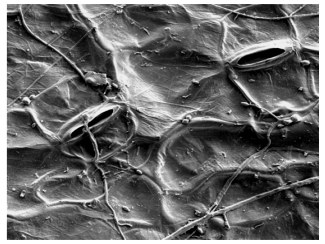
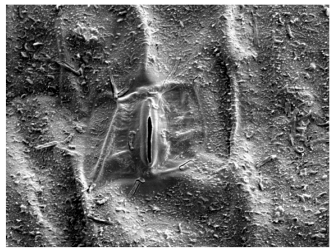
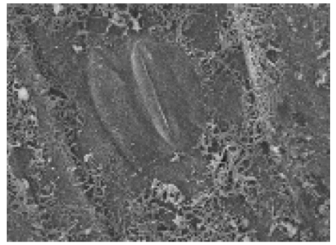
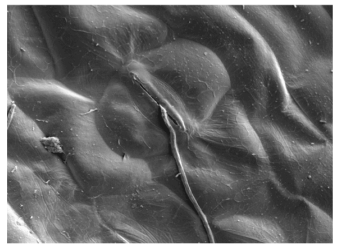
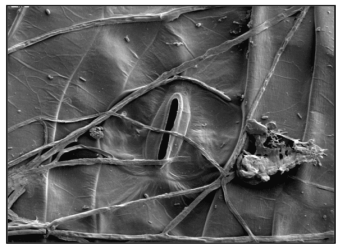
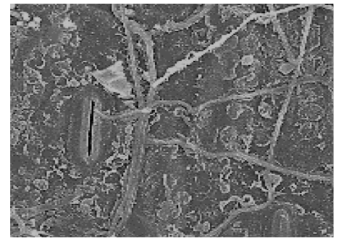
EST= estadío; T= Tiempo

# Pseudotecio liberando ascosporas de *M. fijiensis*




A= ascoporas; Pt= Pseudotecio; h= hifas dentro del tejido

# Seguimiento del hongo *M. fijiensis* en en las variedades de Calcutta 4 y Williams mediante SEM

| EST | E0 | | | E1/2 | E3 |
|---|---|---|---|---|---|
| HPI | 0 | 12 | 24 | 144 | 360 |
| Cal |  |  |  |  |  |
| EST | E0 | | | | E3 |
| HPI | 0 | 12 | 24 | 144 | 360 |
| Will |  |  |  |  |  |

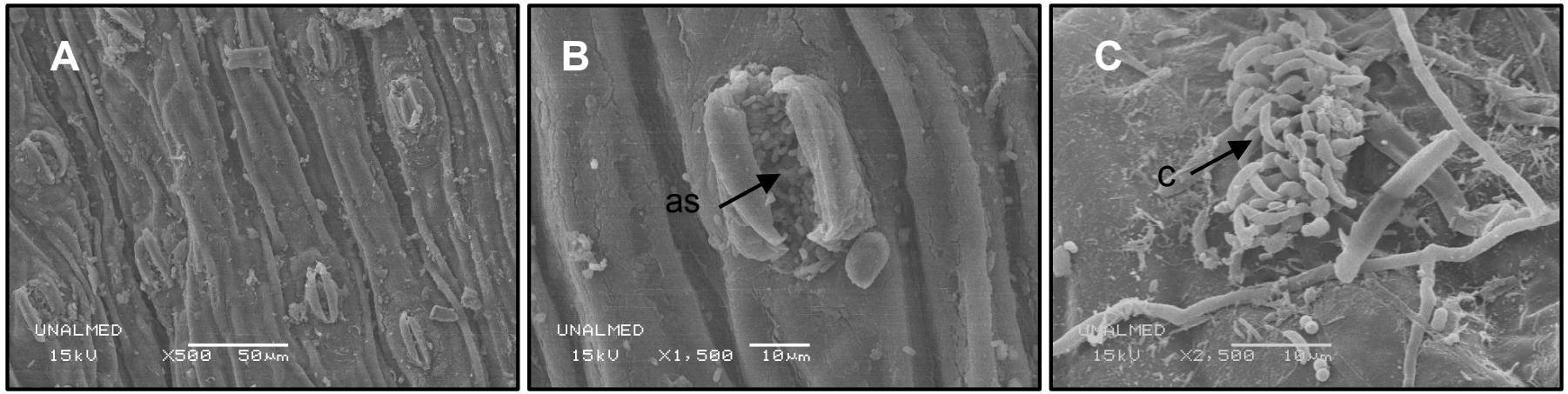
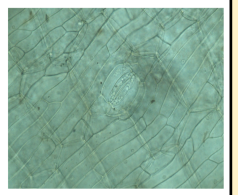
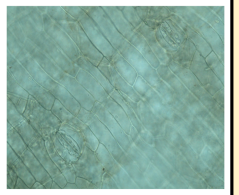
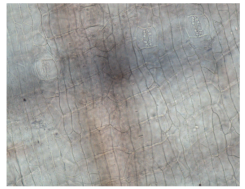
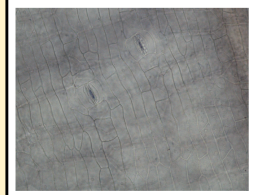
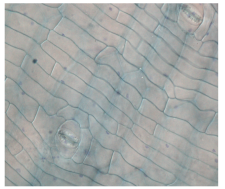
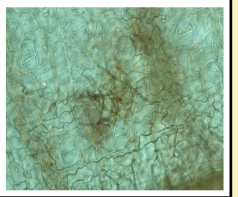
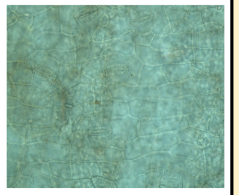
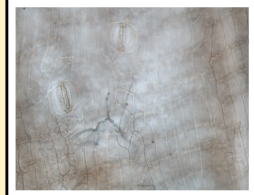
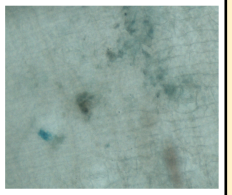
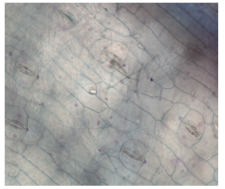
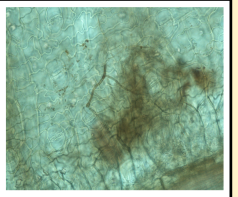
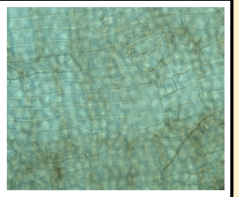
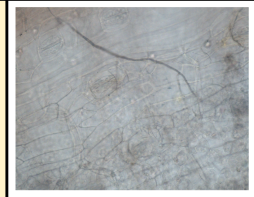
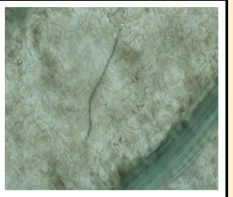
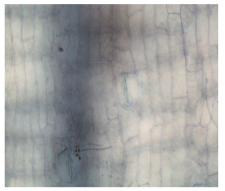
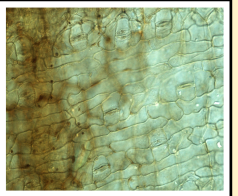
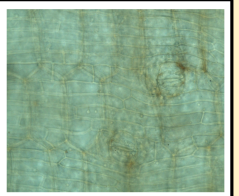
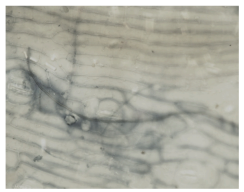
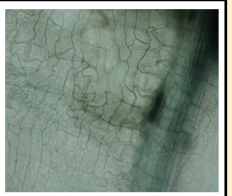
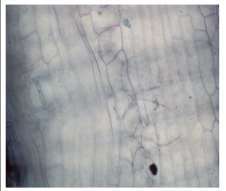
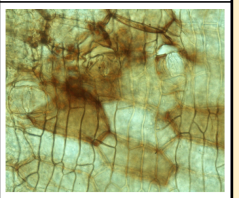
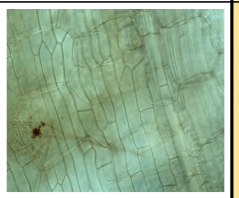
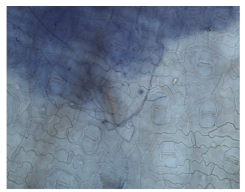
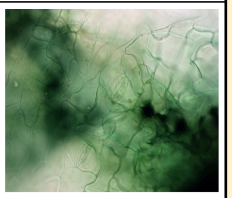
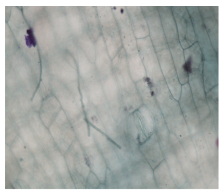
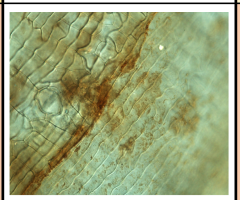
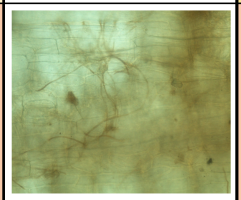
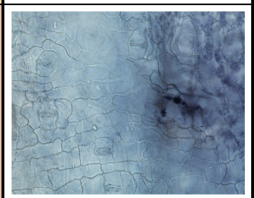
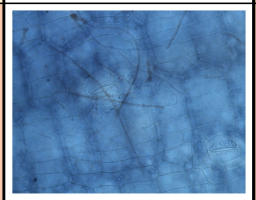
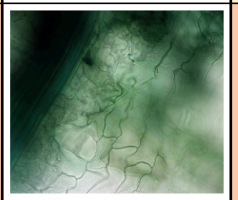
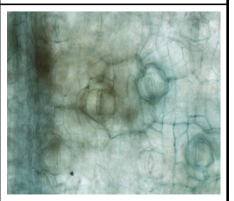
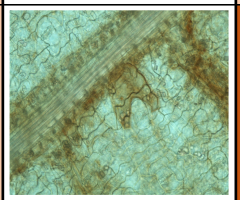
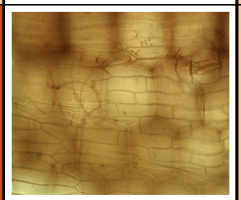
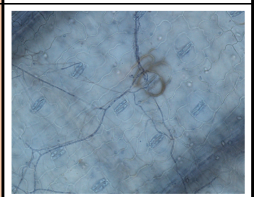
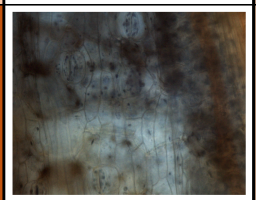
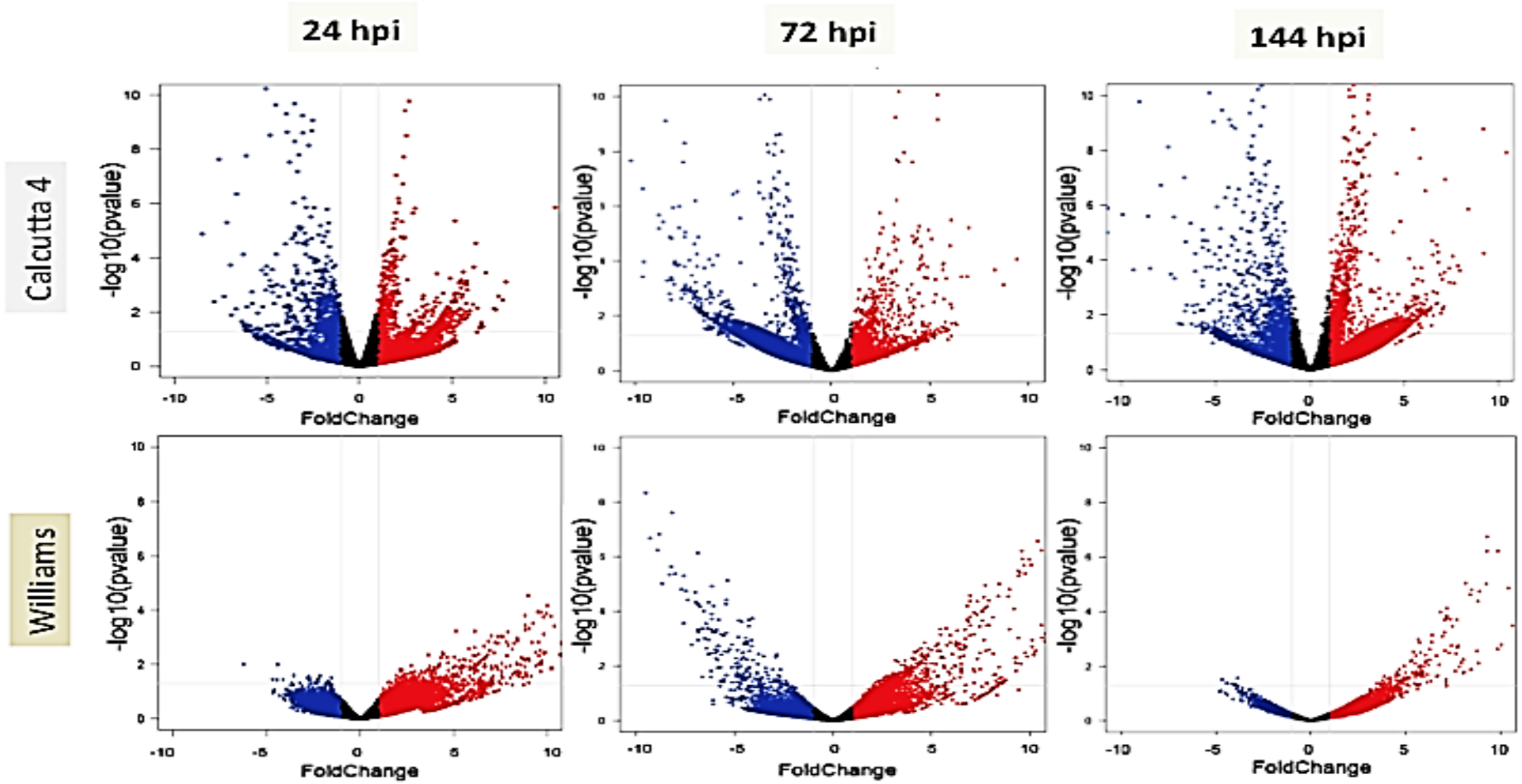



**Figura 2.6.** Fotografías de hojas de Williams durante estadíos avanzados de la Sigatoka Negra. A) Mancha madura durante estadío 5 de la enfermedad; B) Pseudotecio con ascosporas al interior de un estoma; C) Esporodonquios con conidias en mancha joven de la enfermedad. As= ascosporas; c= conidias.

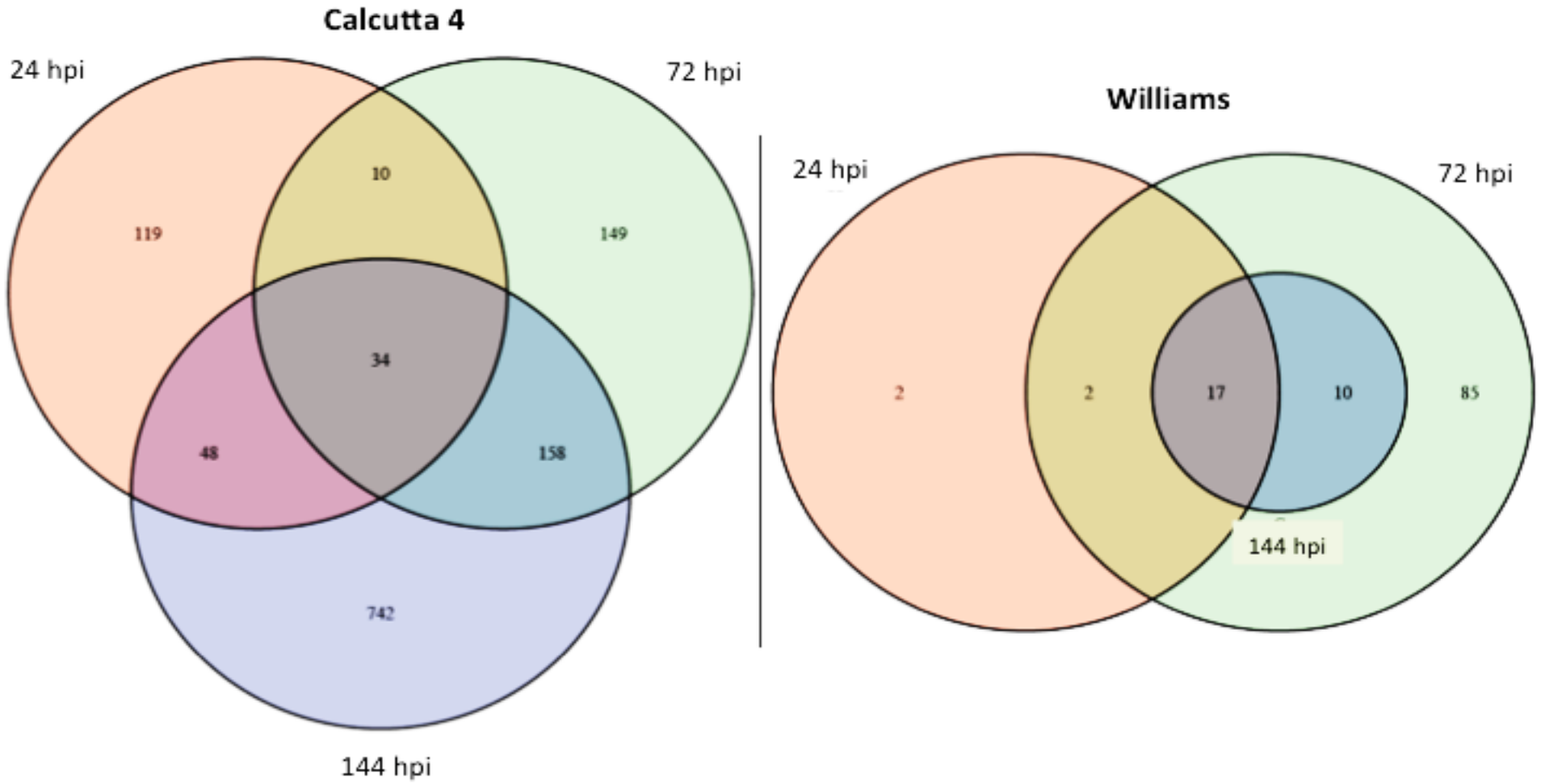
| EST | HPI | Calcutta 4 | EST | HPI | Williams | EST | HPI | Calcutta 4 | EST | HPI | Williams | EST | HPI | Calcutta 4 | EST | HPI | Williams |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E0 | 0 |  | E0 | 0 |  | E0 | 0 |  | E0 | 0 |  | E0 | 0 |  | E0 | 0 |  |
| | 24 |  | | 24 |  | | 24 |  | | 24 |  | | 24 |  | | 24 |  |
| | 48 |  | | 48 |  | | 48 |  | | 48 |  | | 48 |  | | 48 |  |
| | 72 |  | | 72 |  | | 72 |  | | 72 |  | | 72 |  | | 72 |  |
| E1/2 | 144 |  | | 144 |  | E1/2 | 144 |  | | 144 |  | E1/2 | 144 |  | | 144 |  |
| E3 | 360 |  | E3 | 360 |  | E3 | 360 |  | E3 | 360 |  | E3 | 360 |  | E3 | 360 |  |
| E3 | 720 |  | E5 | 720 |  | E3 | 720 |  | E5 | 720 |  | E3 | 720 |  | E5 | 720 |  |

# Perfiles transcriptómicos de genotipos de *Musa acuminata* revelan genes relacionados con la defensa en la interacción con *Mycosphaerella fijiensis*

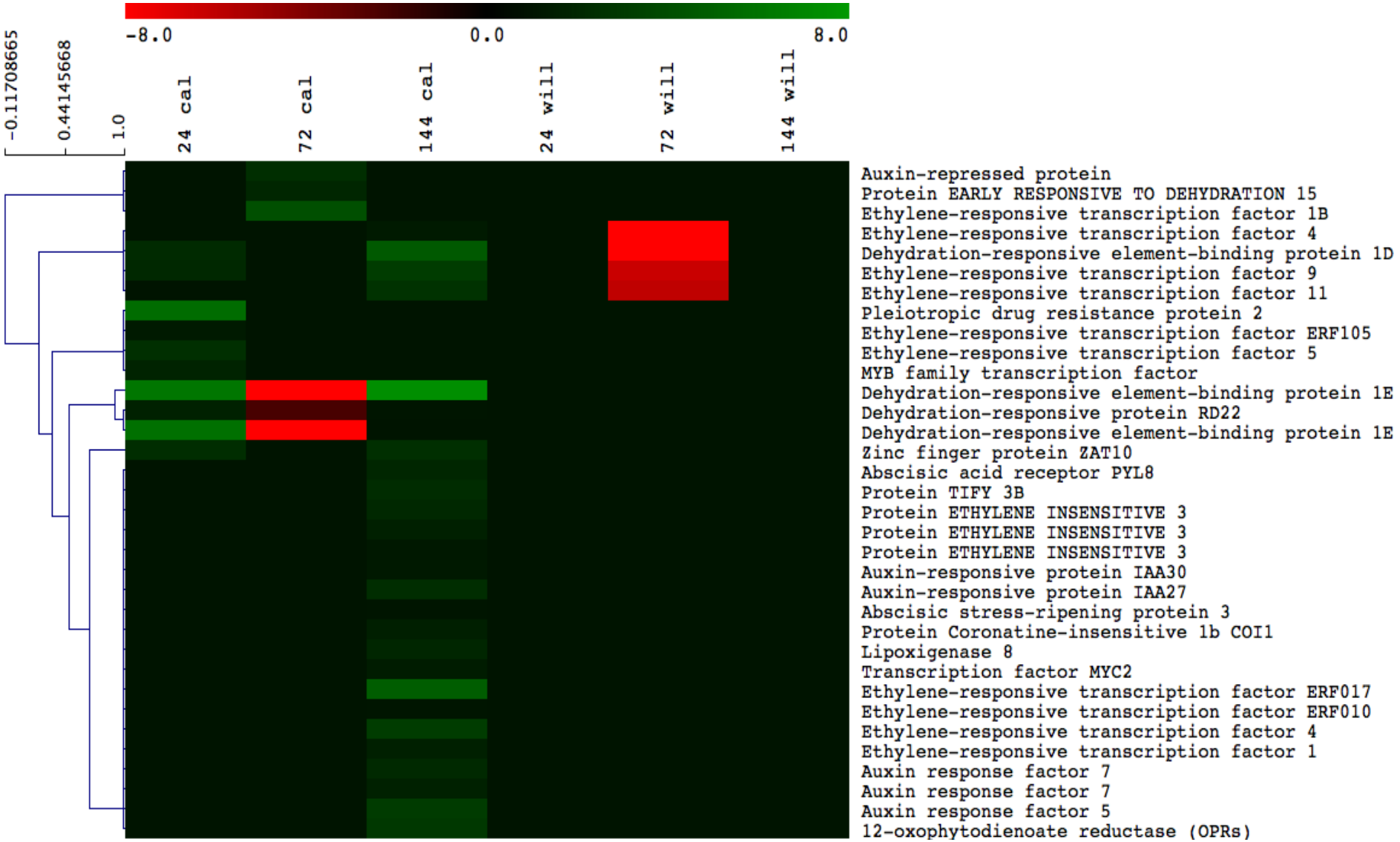
# Volcano plot



# Diagrama de Venn de los genes que se comparten entre los tiempos por variedad

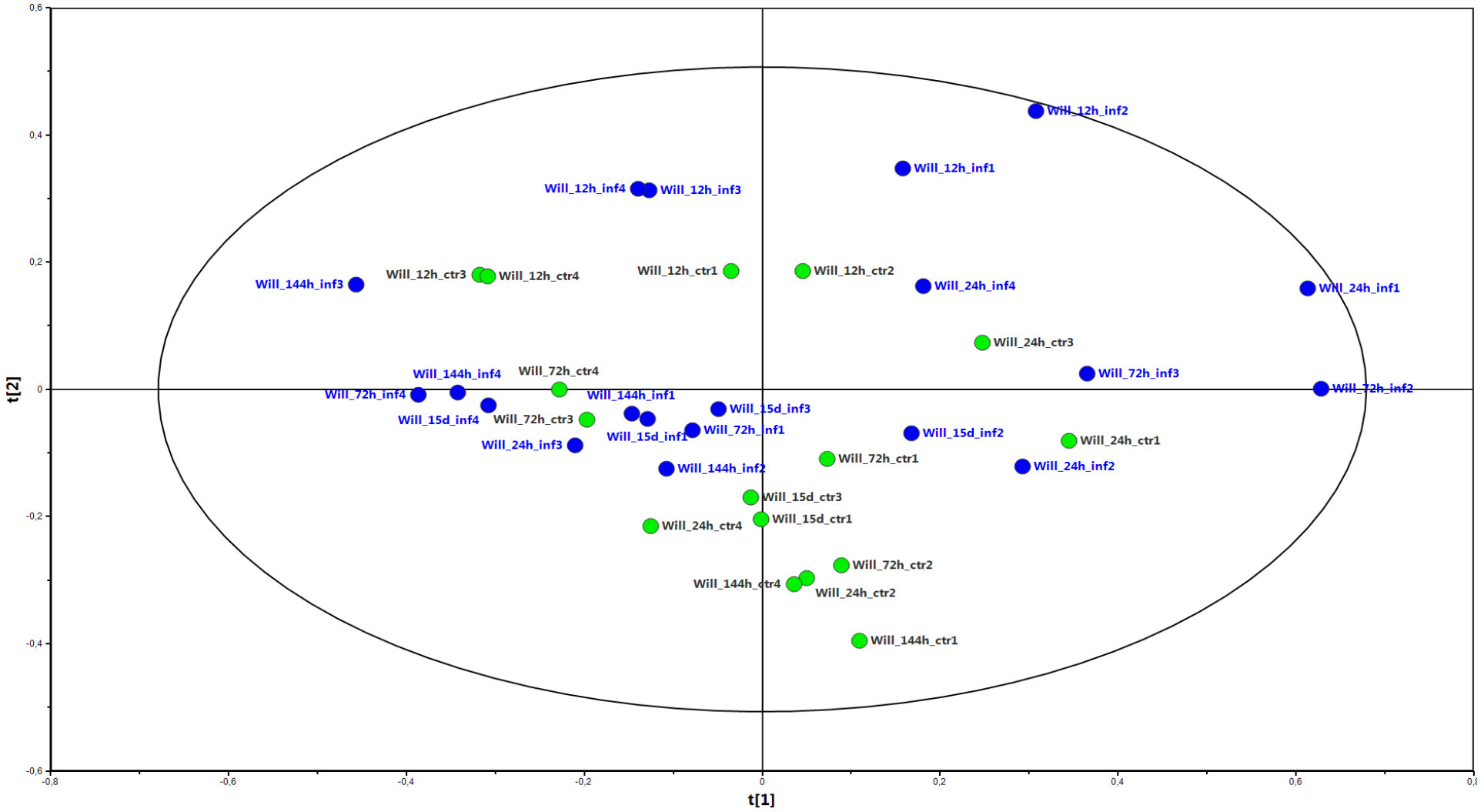


# Transducción de señales hormonales.



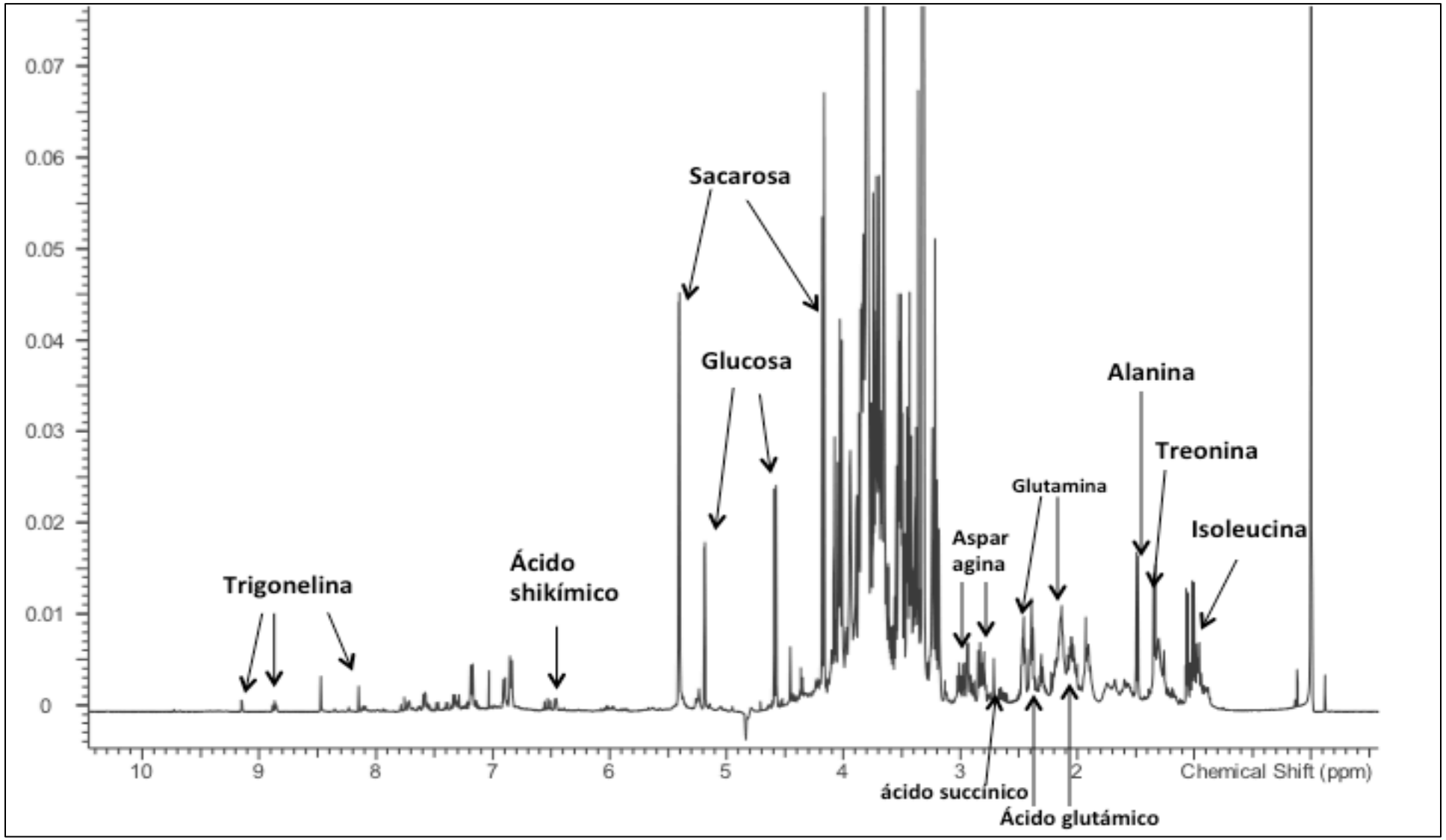
# Perfil metabolico de hojas de banano durante la interacción con el hongo *Mycosphaerella fijiensis*

# Análisis de componentes principales de la región 0.5-9.0 ppm del espectro de $^{1}H$ NMR (500 MHz)

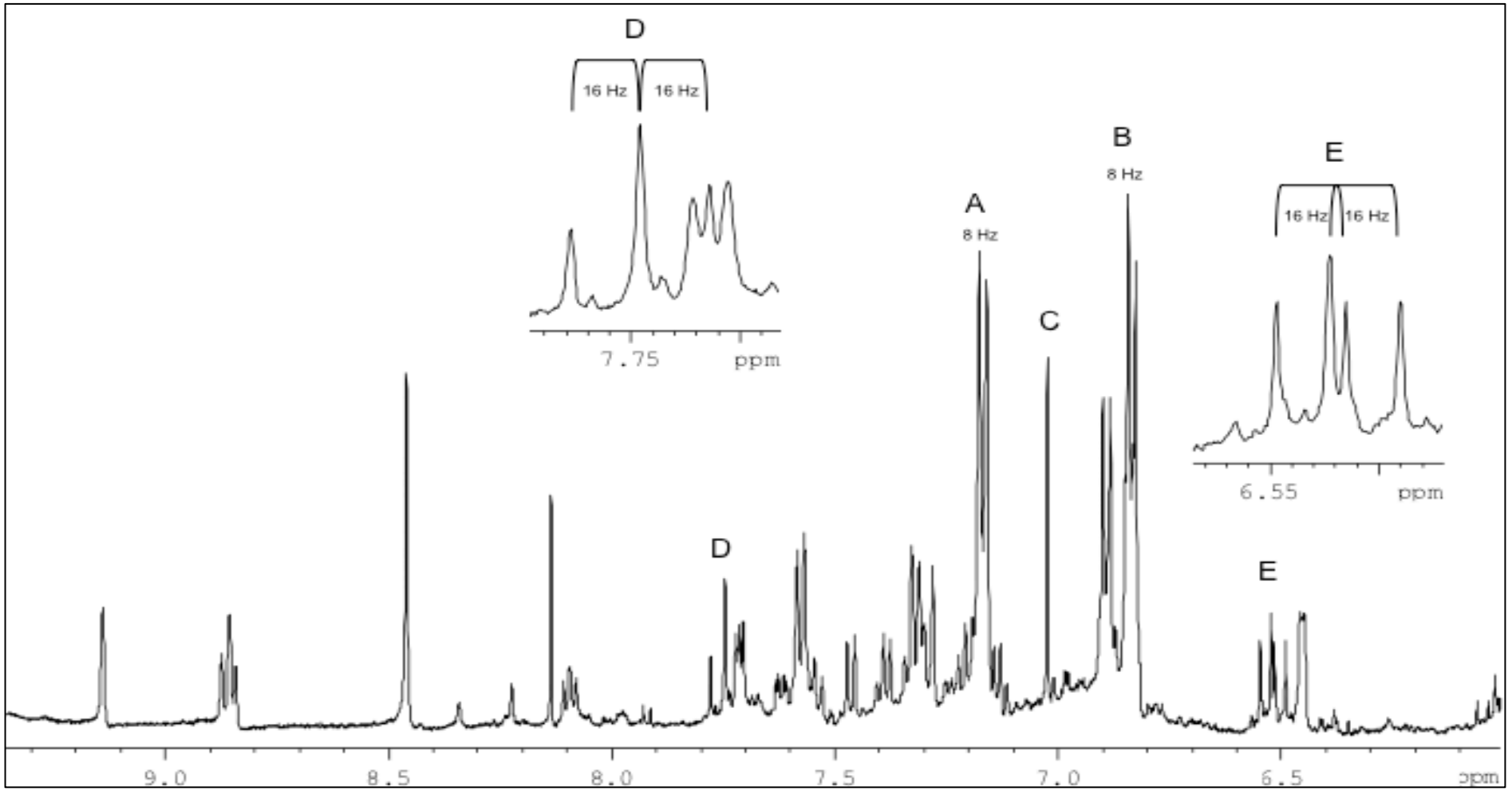



Variedad susceptible Williams

# Metabolitos identificados en los espectros de $^{1}H$ NMR



**Señales relacionadas con fenilfenalenonas (A-E), evidenciando la presencia de dichos compuestos en las muestras**



**Análisis no supervisado de metabolitos diferencialmente expresados en plantas de banano durante la inoculación con *M. fijiensis* por UPLC-ESI-MS**

# Cromatograma de iones totales antes (A) y después (B) de la corrección del tiempo de retención

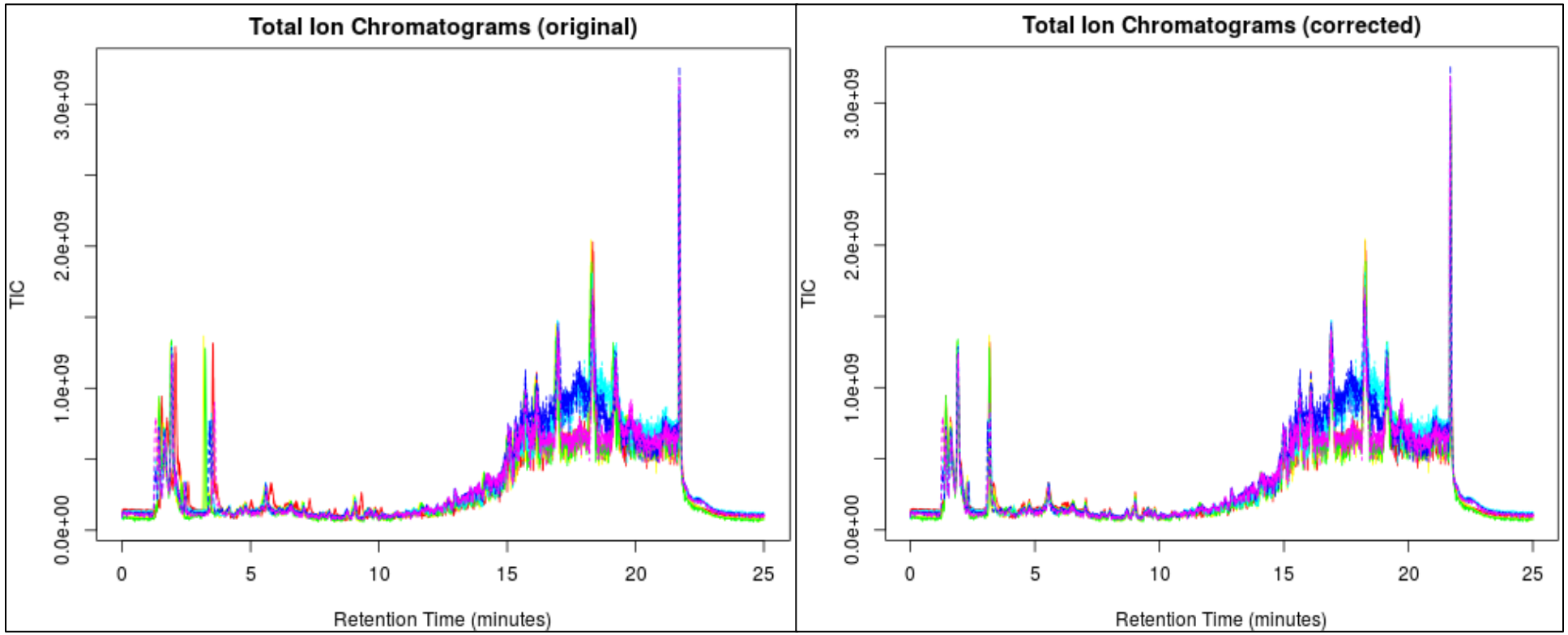
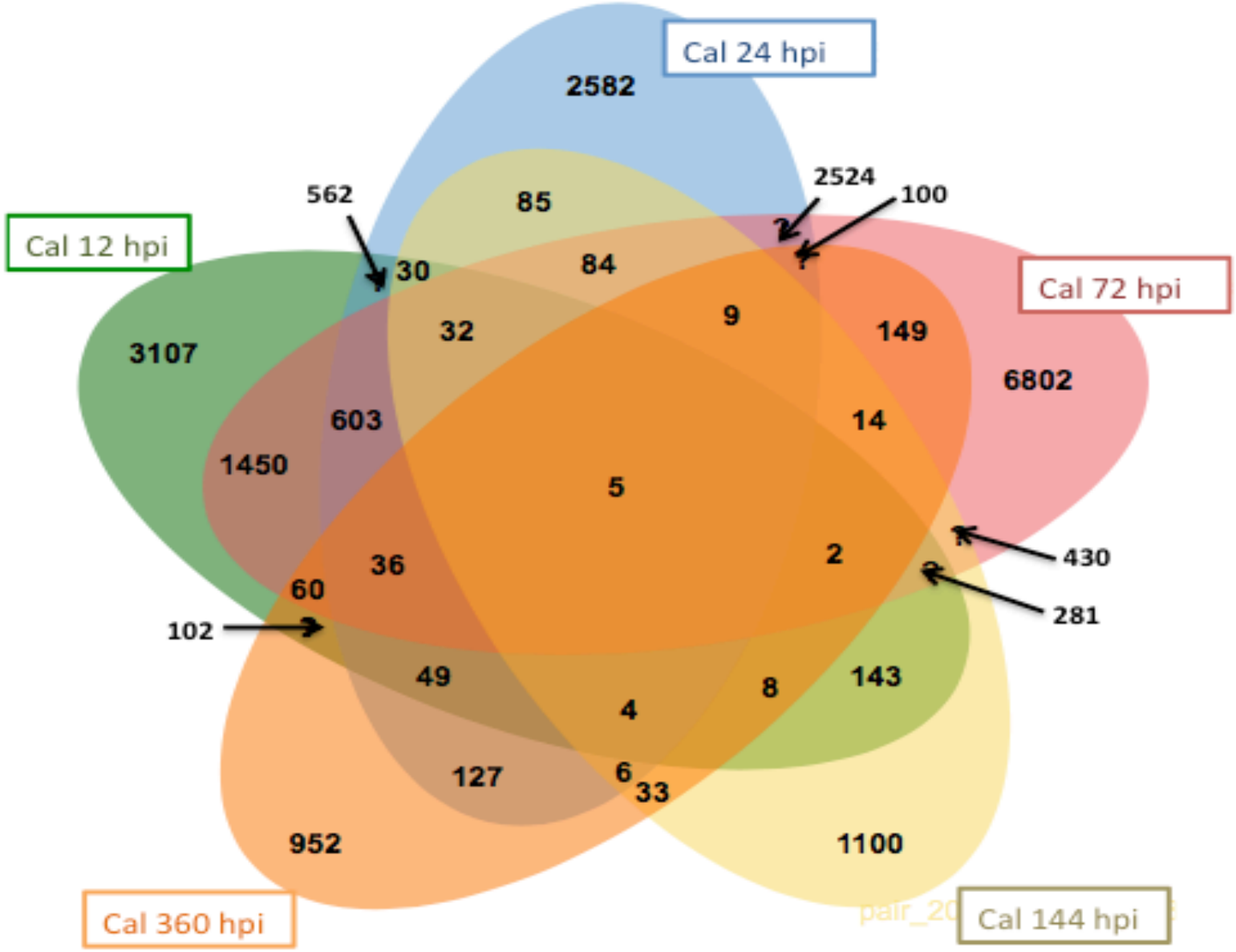


**Diagrama de Venn para comparar las características identificadas por cada tiempo en la variedad Calcutta 4**



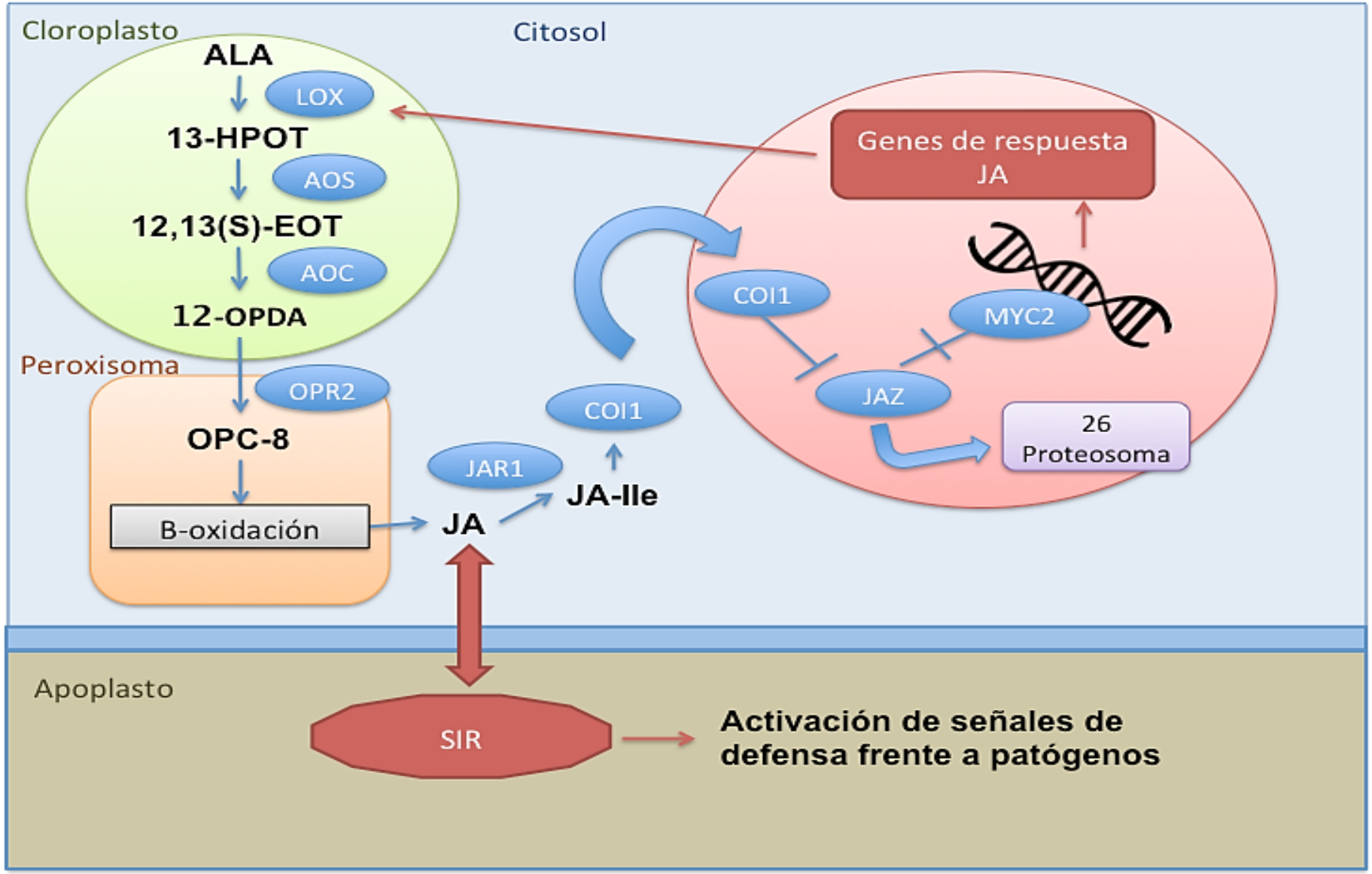
# Principales 30 rutas identificadas para Calcutta 4 a 72 hpi con el hongo *M. fijiensis*

| | | | | | |
|---|---|---|---|---|---|
| sucrose biosynthesis II | "SOBRE" | primario | 5 | 10 | 9.7e-3 |
| sucrose degradation II (sucrose synthase) | "SOBRE" | primario | 4 | 8 | 4.0e-3 |
| Calvin-Benson-Bassham cycle | "SOBRE" | primario | 4 | 12 | 2.1e-2 |
| glycolysis I (from glucose 6-phosphate) | "SOBRE" | primario | 4 | 11 | 1.5e-1 |
| gluconeogenesis I | "SOBRE" | primario | 4 | 11 | 1.5e-1 |
| trehalose degradation II (trehalase) | "SOBRE" | primario | 3 | 4 | 5.9e-4 |
| guanosine nucleotides degradation I | "SOBRE" | primario | 3 | 5 | 9.4e-4 |
| pentose phosphate pathway (non-oxidative branch) | "SOBRE" | primario | 3 | 6 | 1.5e-3 |
| mannitol biosynthesis | "SOBRE" | primario | 3 | 4 | 9.6e-3 |
| Rubisco shunt | "SOBRE" | primario | 3 | 11 | 1.5e-2 |
| chlorophyllide abiosynthesis I (aerobic, light-dependent) | "SUB" | primario | 3 | 13 | 2.2e-1 |

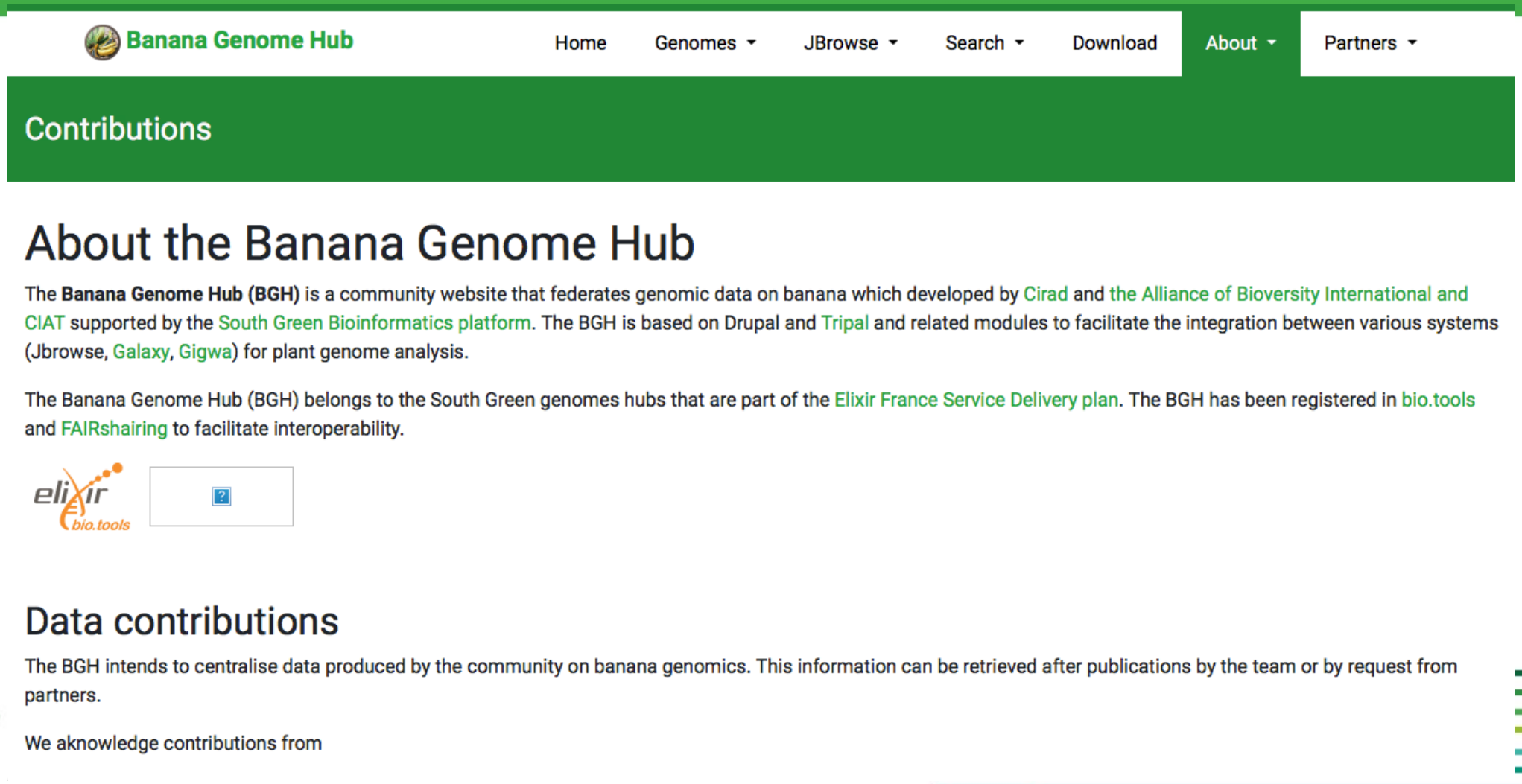
# Principales 30 rutas identificadas para Calcutta 4 a 72 hpi con el hongo *M. fijiensis*

| Ruta | Regulación | tipo de metabolismo | Metabolitos Putativos | Metabolitos Ruta | Valor p |
|---|---|---|---|---|---|
| flavonoid biosynthesis | "SOBRE" | secundario | 8 | 13 | 3.3e-4 |
| kaempferol glycoside biosynthesis | "SOBRE" | secundario | 5 | 14 | 9.5e-4 |
| tetrahydrofolate biosynthesis II | "SOBRE" | secundario | 4 | 18 | 1.8e-2 |
| luteolin biosynthesis | "SOBRE" | secundario | 3 | 6 | 1.5e-3 |
| rosmarinic acid biosynthesis II | "SOBRE" | secundario | 3 | 8 | 4.0e-3 |
| coumarin biosynthesis (via 2-coumarate) | "SOBRE" | secundario | 3 | 8 | 4.0e-3 |
| flavonol biosynthesis | "SOBRE" | secundario | 3 | 8 | 4.0e-3 |
| monolignol glucosides biosynthesis | "SOBRE" | secundario | 3 | 10 | 9.7e-3 |
| traumatin and (Z)-3-hexen-1-yl acetate biosynthesis | "SOBRE" | secundario | 3 | 10 | 9.7e-3 |
| carotenoid cleavage | "SOBRE" | secundario | 3 | 21 | 1.9e-1 |
| phenylpropanoid biosynthesis | "SUB" | secundario | 3 | 17 | 3.8e-1 |

**Reconstrucción de la ruta putativa de biosíntesis y transducción de señales de JA en la variedad Calcutta 4 durante la interacción con *Mycosphaerella fijiensis***

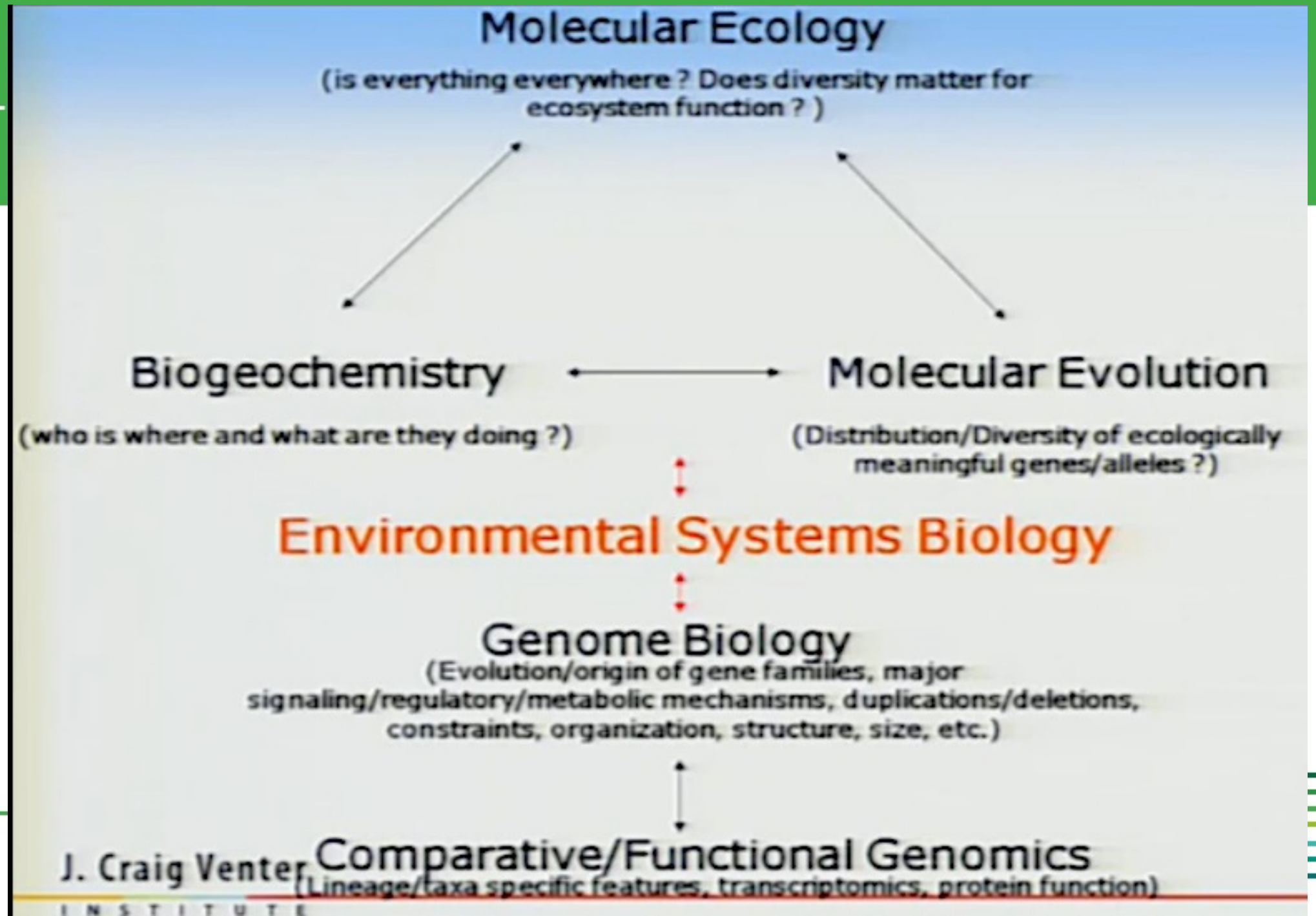


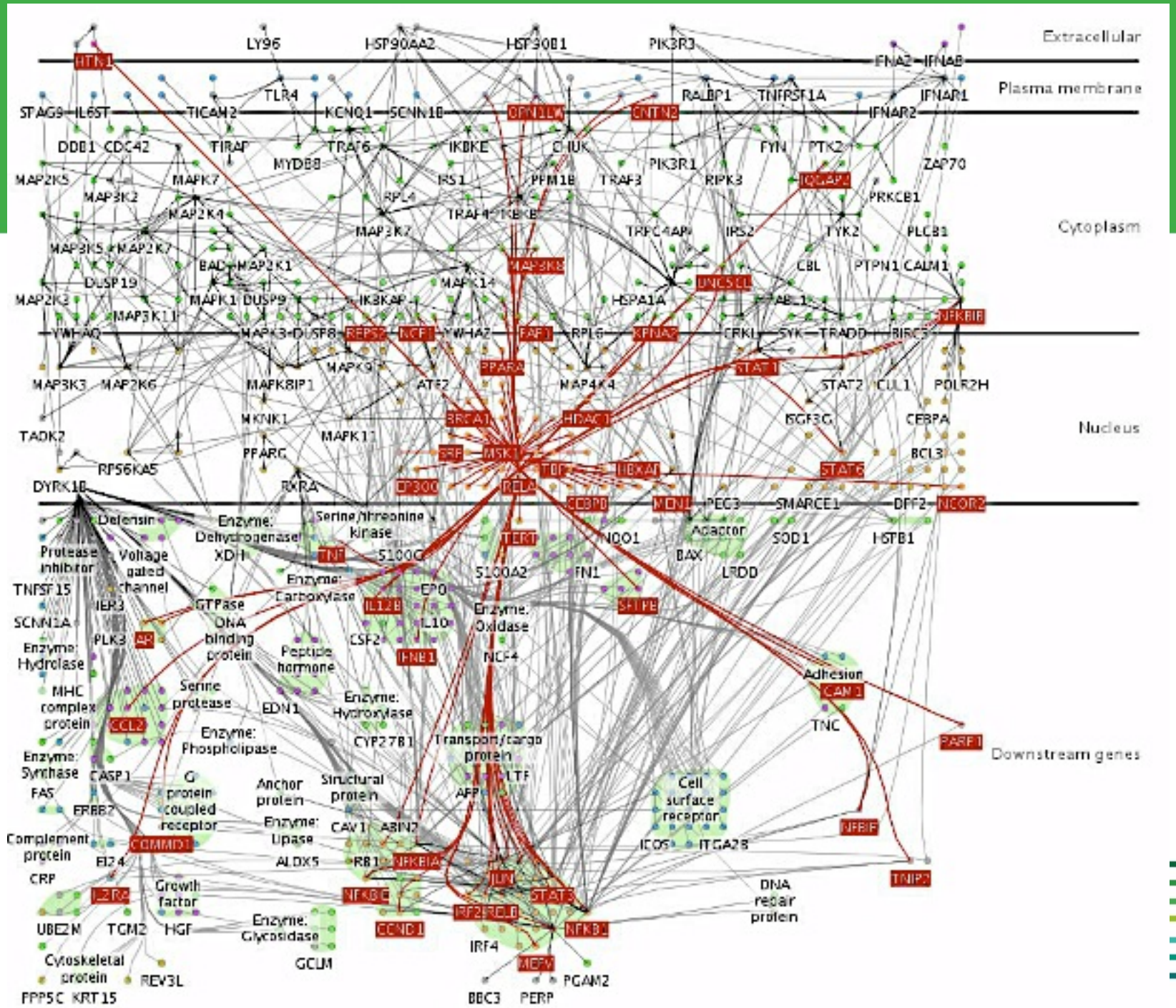
# Big data para la Ciencia



Ellos encuentran mas de 1800 especies de microbios incluyendo 150 especies nuevas de bacterias y mas de 1.2 millones de nuevos genes.

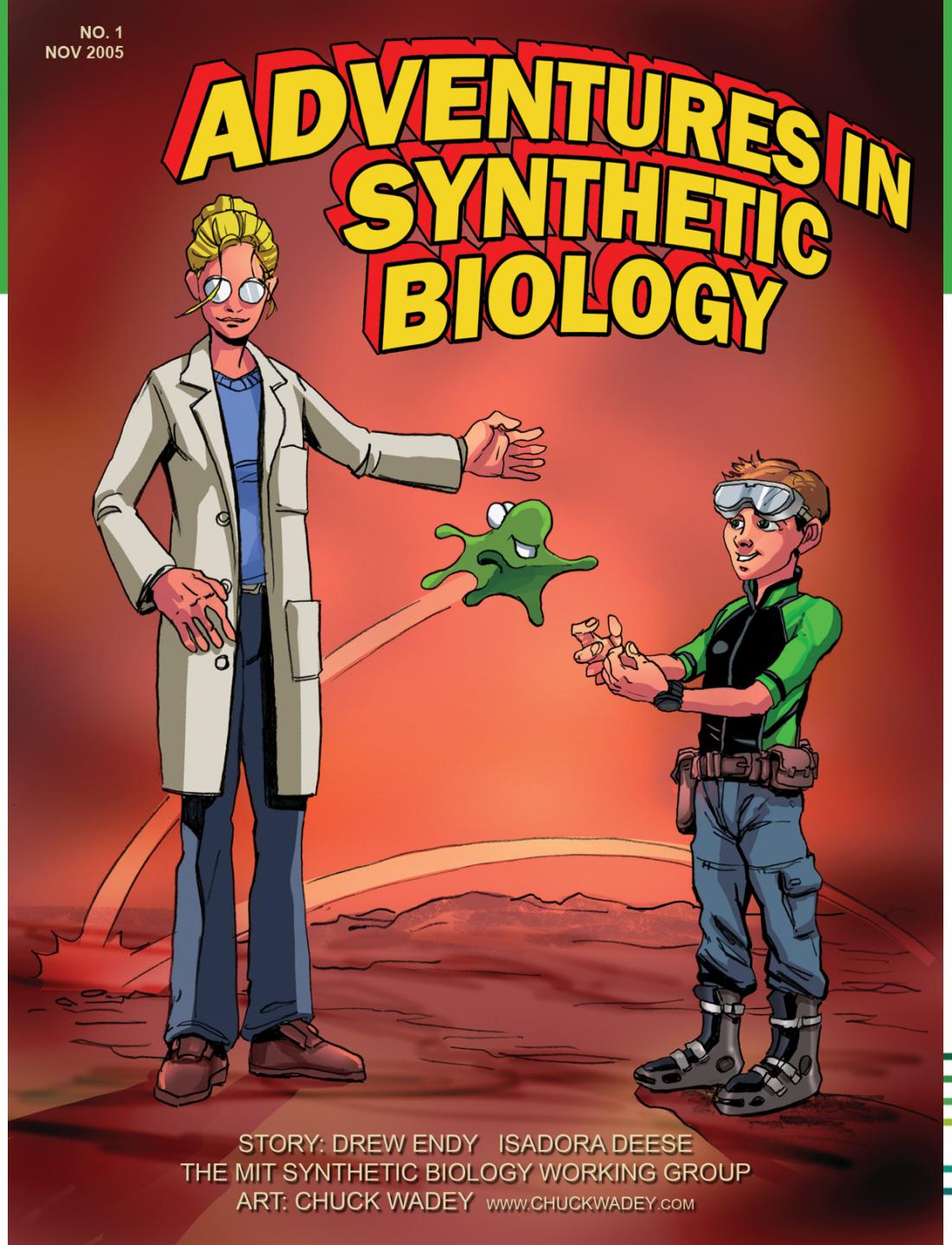
Collecting sea water samples from the Sargasso Sea for the whole genome shotgun sequencing of microbial populations. (Photo supplied by The Center for the Advancement of Genomics.)


Molecular Ecology
(is everything everywhere ? Does diversity matter for ecosystem function ? )
Biogeochemistry
(who is where and what are they doing ?)
Molecular Evolution
(Distribution/Diversity of ecologically meaningful genes/alleles ?)
Environmental Systems Biology
Genome Biology
(Evolution/origin of gene families, major signaling/regulatory/metabolic mechanisms, duplications/deletions, constraints, organization, structure, size, etc.)
Comparative/Functional Genomics
(Lineage/taxa specific features, transcriptomics, protein function)
J. Craig Venter
INSTITUTE

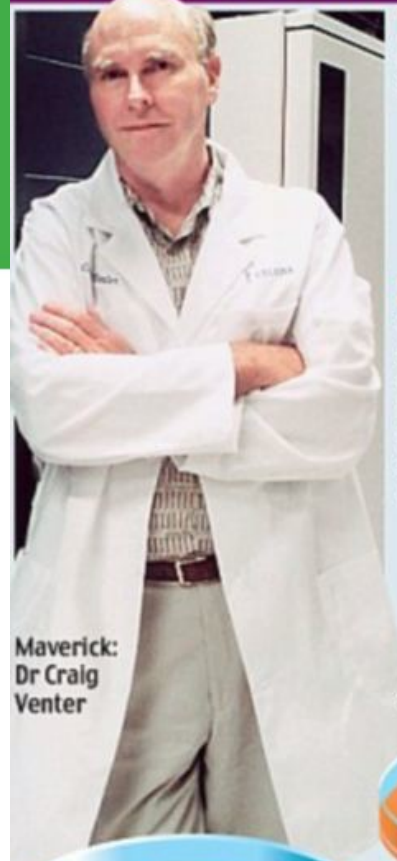

Extracellular
Plasma membrane
Cytoplasm
Nucleus
Downstream genes
HTN1
LY96
HSP90AA2
HSP90B1
PIK3R3
IFNA2
IFNAB
TLR4
RALBP1
TNFRSF1A
IFNAR1
IFNAR2
STAG3
IL6ST
TICAM2
KCNQ1
SCNN1B
CNTN2
DDB1
CDC42
TIRAP
TRAF6
IKBKE
CHUK
FYN
PTK2
ZAP70
MAP2K5
MAPK7
MYD88
IRS1
PPM1B
TRAF3
RIPK3
PRKCB1
MAP3K2
MAP2K4
RPL4
TRAF4
TBK1
MAP3K7
TRPC4AP
IRS2
TYK2
PLCB1
MAP3K5
MAP2K7
MAP2K8
CBL
PTPN1
CALM1
BAD
MAP2K1
DUSP19
MAPK14
UNC5CL
MAP2K3
MAPK1
DUSP9
IKBKAP
HSPA1A
ABL1
NFKBIB
MAP3K11
YWHAQ
MAPK3
DUSP8
REPS2
NCF1
YWHAZ
FAF1
RPL6
KPNA2
CRKL
SYK
TRADD
BIRC3
MAPK9
PPARA
STAT1
MAP3K3
MAP2K6
MAPK8IP1
ATF2
MAP4K4
STAT2
CUL1
POLR2H
MKNK1
BRCA1
HDAC1
IGF3G
CEBPA
TADK2
MAPK11
PPARG
MSK1
BCL3
RPS6KA5
TOP
HDAC7
STAT6
DYRK1B
RXRA
EP300
RELA
SCBP0
MCM1
PEG3
SMARCE1
DFF2
NCOR2
Defensin
Enzyme: Dehydrogenase
Serine/threonine kinase
Protease inhibitor
Voltage gated channel
XDH
TNF
S100G
TERT
NQO1
Adaptor
SOD1
HSTB1
TNFSF15
IER3
GTPase
Enzyme: Carboxylase
S100A2
FN1
BAX
LRDD
IL12B
EPO
SH2B
SCNN1A
PLK3
AP
DNA binding protein
IL10
Enzyme: Oxidase
Enzyme: Hydrolase
CSF2
Peptide hormone
HNB1
NCF4
Adhesion
CAN1
MHC complex protein
Serine protease
EDN1
Enzyme: Hydroxylase
TNC
CCL2
Enzyme: Phosphatase
CYP27B1
Transport/cargo protein
PARP1
Enzyme: Synthase
CASP1
Anchor protein
Structural protein
LTF
FAS
G protein coupled receptor
Cell surface receptor
ERBB2
CAV1
ABIN2
Enzyme: Lipase
NFRIF
Complement protein
COMMD1
ITGA2B
EIF4
ALOX5
RB1
NFKBIA
CRP
JUN
TNIP2
Growth factor
NFKBIE
STAT3
DNA repair protein
IL2RG
IRF2BPL
UBE2M
TCM2
HGF
Enzyme: Glycosidase
CCND1
NFKB1
IRF4
Cytoskeletal protein
GCLM
MEFV
REV3L
PGAM2
PPP5C
KRT15
BBC3
PERP

# ¿Cual es el siguiente paso?

¿Como puedo utilizar toda esta información?


NO. 1
NOV 2005
ADVENTURES IN SYNTHETIC BIOLOGY
STORY: DREW ENDY ISADORA DEESE
THE MIT SYNTHETIC BIOLOGY WORKING GROUP
ART: CHUCK WADEY WWW.CHUCKWADEY.COM

# HOW TO MAKE ARTIFICIAL LIFE



Maverick: Dr Craig Venter

**1** Entire DNA of Mycoplasma mycoides, a bug that usually infects goats, is decoded.



Synthetic DNA code

**2** Researchers buy fragments of DNA from a mail order catalogue. Each of the four bottles of chemicals contains a section of the code.

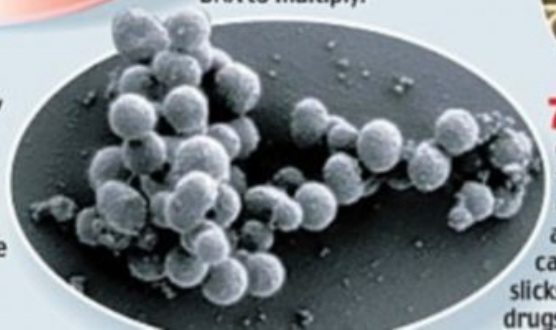
**3** The fragments are put into yeast, which 'stitches' them together, gradually building a synthetic copy of the original DNA.

Artificial DNA

Natural DNA

**4** The artificial DNA is put into a recipient bacterium, which then grows and divides, creating two daughter cells, one with the artificial DNA and one with the natural DNA.

**5** Antibiotics in the petri dish kill the bacterium with the natural DNA, leaving the one with the synthetic DNA to multiply.



**6** Within just a few hours, all traces of the recipient bug are wiped out and bugs with artificial DNA thrive. New life has been created.

Graphic by John Lawson

**7** Possible uses are bugs capable of producing clea fuels and sucking carbon dioxide out of the atmosphere. Also microbe capable of mopping up oil slicks (above) or generating drugs, including the flu vaccine

# Big data para la Ciencia

American College of Neuropsychopharmacology

www.nature.com/npp

ARTICLE OPEN

Check for updates

## Genome-wide meta-analysis of alcohol use disorder in East Asians

Hang Zhou[1,2], Rasmon Kalayasiri[3,4,5], Yan Sun[6], Yaira Z. Nuñez[1,2], Hong-Wen Deng[7], Xiang-Ding Chen[8], Amy C. Justice[9,10], Henry R. Kranzler[11,12], Suhua Chang[13], Lin Lu[6,13], Jie Shi[6], Kittipong Sanichwankul[14], Apiwat Mutirangura[5], Robert T. Malison[1,16] and Joel Gelernter[1,2,15]✉

© The Author(s) 2022

Alcohol use disorder (AUD) is a leading cause of death and disability worldwide. Genome-wide association studies (GWAS) have identified ~30 AUD risk genes in European populations, but many fewer in East Asians. We conducted GWAS and genome-wide meta-analysis of AUD in 13,551 subjects with East Asian ancestry, using published summary data and newly genotyped data from five cohorts: (1) electronic health record (EHR)-diagnosed AUD in the Million Veteran Program (MVP) sample; (2) DSM-IV diagnosed alcohol dependence (AD) in a Han Chinese–GSA (array) cohort; (3) AD in a Han Chinese–Cyto (array) cohort; and (4) two AD Thai cohorts. The MVP and Thai samples included newly genotyped subjects from ongoing recruitment. In total, 2254 cases and 11,297 controls were analyzed. An AUD polygenic risk score was analyzed in an independent sample with 4464 East Asians (Genetic Epidemiology Research in Adult Health and Aging (GERA)). Phenotypes from survey data and ICD-9-CM diagnoses were tested for association with the AUD PRS. Two risk loci were detected: the well-known functional variant rs1229984 in *ADH1B* and rs3782886 in *BRAP* (near the *ALDH2* gene locus) are the lead variants. AUD PRS was significantly associated with days per week of alcohol consumption (beta = 0.43, SE = 0.067, $p = 2.47 \times 10^{-10}$) and nominally associated with pack years of smoking (beta = 0.09, SE = 0.05, $p = 4.52 \times 10^{-2}$) and ever vs. never smoking (beta = 0.06, SE = 0.02, $p = 1.14 \times 10^{-2}$). This is the largest GWAS of AUD in East Asians to date. Building on previous findings, we were able to analyze pleiotropy, but did not identify any new risk regions,

# Big data para la Ciencia

ARTICLE

https://doi.org/10.1038/s41586-018-0043-0

# Renewing Felsenstein's phylogenetic bootstrap in the era of big data

F. Lemoine[1,2], J.-B. Domelevo Entfellner[3,4], E. Wilkinson[5], D. Correia[1], M. Dávila Felipe[1], T. De Oliveira[5,6] & O. Gascuel[1,7]*

Felsenstein's application of the bootstrap method to evolutionary tr
time. The bootstrap method, which is based on resampling and replic
of phylogenetic inferences. However, increasing numbers of sequenc
and phylogenies based on hundreds or thousands of taxa are becoming
bootstrap tends to yield very low supports, especially on deep branches
bootstrap in which the presence of inferred branches in replications is
than the binary presence or absence index used in Felsenstein's origi
do not induce falsely supported branches. The application of our met
reveals their phylogenetic signals, whereas Felsenstein's bootstrap fa

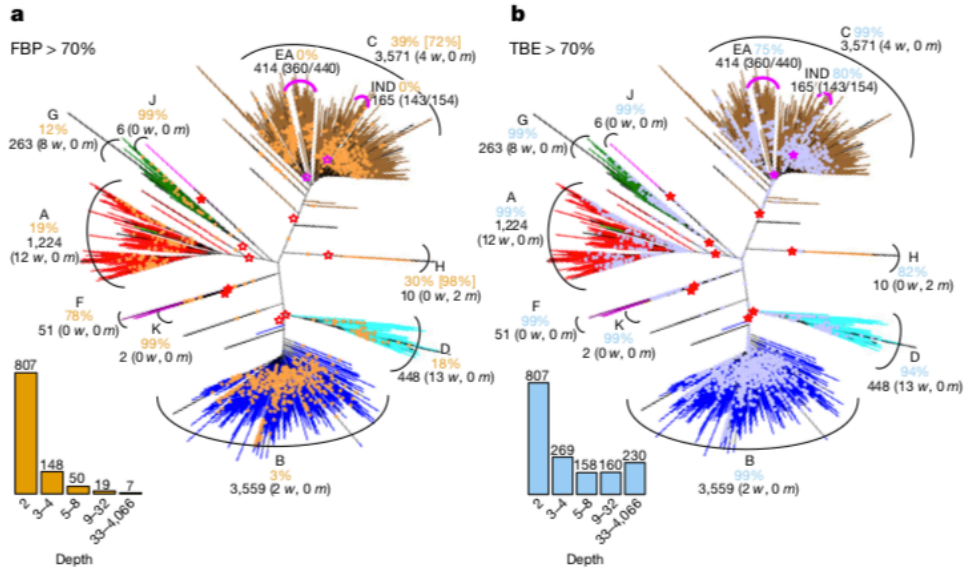
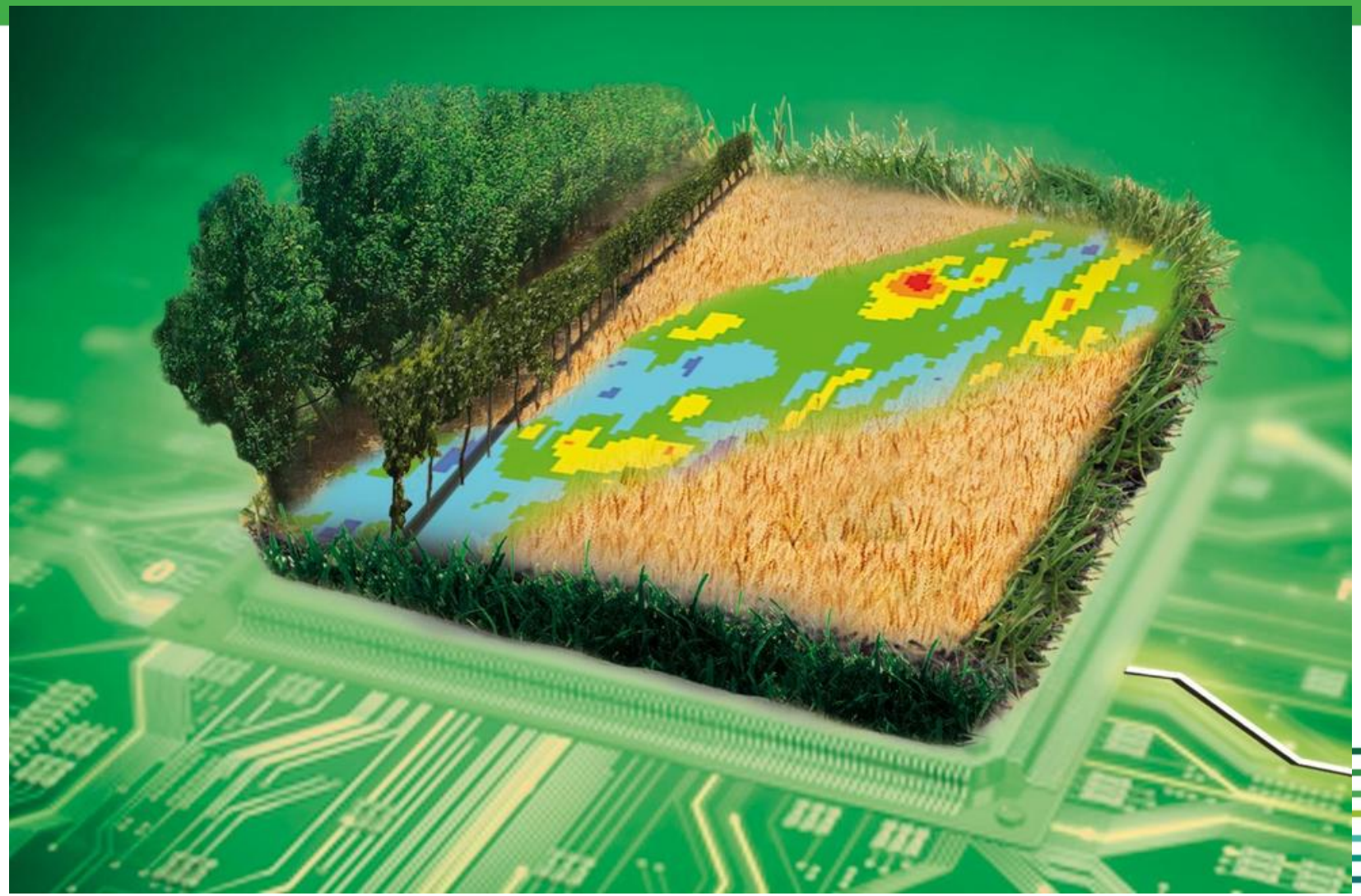



**Fig. 1 | FBP and TBE bootstrap supports on the same phylogeny with 9,147 HIV-1M *pol* sequences. a**, FBP. **b**, TBE. Subtypes are colourized; supports larger than 70%, which are shown in square brackets. The same approach is applied to the C sub-epidemics in India (IND) and East Africa

# Big data para la Ciencia



# Big data para la Ciencia

Article | Open Access | Published: 13 October 2021

## From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling

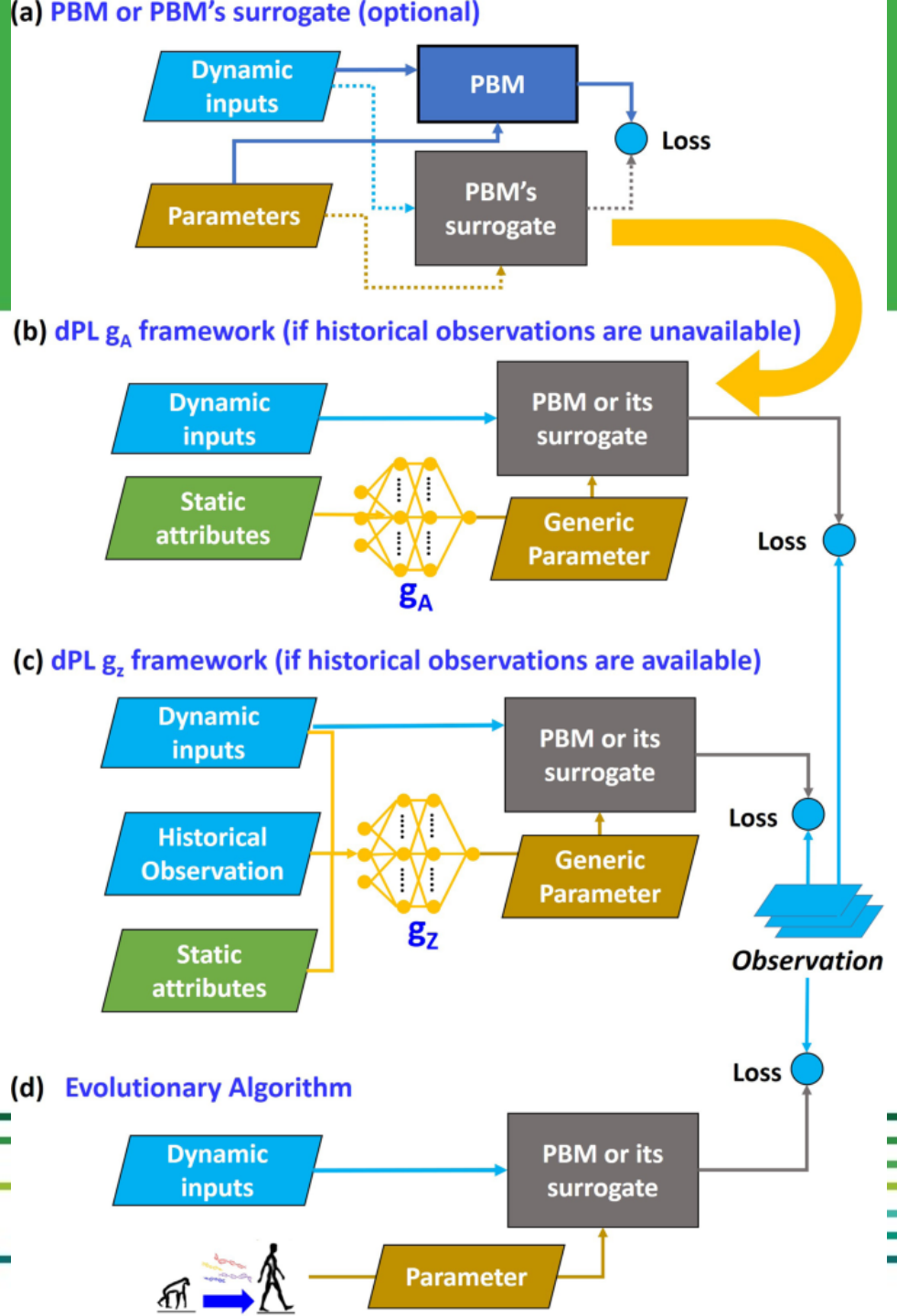
Wen-Ping Tsai, Dapeng Feng, Ming Pan, Hylke Beck, Kathryn Lawson, Yuan Yang, Jiangtao Liu & Chaopeng Shen

*Nature Communications* **12**, Article number: 5988 (2021) | Cite this article

**8313** Accesses | **3** Citations | **64** Altmetric | Metrics

### Abstract

The behaviors and skills of models in many geosciences (e.g., hydrology and ecosystem sciences) strongly depend on spatially-varying parameters that need calibration. A well-calibrated model can reasonably propagate information from observations to unobserved variables via model physics, but traditional calibration is highly inefficient and results in non-unique solutions. Here we propose a novel differentiable parameter learning (dPL) framework that efficiently learns a global mapping between inputs (and optionally responses) and parameters. Crucially, dPL exhibits beneficial scaling curves not previously demonstrated to geoscientists: as training data increases, dPL achieves better performance, more physical coherence, and better generalizability (across space and uncalibrated



# Perspectivas

Muchas Gracias por su atención ….


# UNIVERSIDAD DE ANTIOQUIA

@UdeA @UdeA @universidaddeantioquia