Presentación de la base de datos asociada al proyecto "***Diagnóstico de la biodiversidad genética de razas y variedades de maíz nativo para la toma de decisiones y la evaluación de programas de conservación***"

Octavio Martínez de la Vega

Cibiogem

30 de Julio de 2015

Utilizando
GenoMaíz

(http://computational.biology.langebio.cinvestav.mx/GenoMaiz/)

Octavio Martínez de la Vega
omartine@langebio.cinvestav.mx

30 de Julio de 2015
CIBIOGEM

Vínculos en la página principal

- Resumen
- Colaboradores
- Responsables Técnicos
- Fotos de Semillas
- Consultas
- Mapas con Accesiones
- Documentos.

¿Base de datos (relacionales)?

- Colección de ***tablas*** (renglones x columnas)
- Podemos conceptualizar:
  - Columnas = Variables
  - Renglones = Instancias (***datos*** *[números u otros caracteres…]*)
- Las tablas pueden estar relacionadas entre sí por ***variables*** comunes
- Pueden hacerse ***búsquedas*** rápidas y eficientes (*data mining*)

El lenguaje SQL

- SQL – Por las siglas en inglés de "*Lenguaje de búsquedas estructuradas*"
- Para comunicarnos con la base de datos decimos las palabras mágicas:
- **select** <¿qué?> **from** <tabla(s)> **where** <condiciones>
- Y ¡listo!
- Sintaxis del comando "select" AQUÍ

El lenguaje SQL

- En la interfase en red solamente podemos hacer enunciados “select”
- Sin embargo, en dichos enunciados podemos contar, clasificar, realizar operaciones aritméticas y estadísticas, comparar, limitar, …
- En resumen:
  - Si la información existe en la base de datos, es posible extraerla de muchas maneras distintas. A esto se le conoce como “Minería de datos” y puede resultar en descubrimientos interesantes.

¿qué tipo de consultas podemos hacer?

- Ver si algún término “interesante” existe en la base de datos
- Podemos contar las accesiones agrupándolas por algún criterio (localidad, color, etc.)
- Ver que accesiones tienen un determinado “genotipo” alelo para un determinado marcador
- Determinar la frecuencia de cada alelo por tipo…
- Etc, etc, …

Aprendiendo por ejemplos

- Revisaremos muy rápidamente ejemplos del lenguaje SQL aplicado a la BD *GenoMaíz*
- Para hacer búsquedas razonables es necesario conocer la estructura de las tablas en la BD
  - Podemos consultar la estructura de la(s) tabla(s) para darle formato a nuestra búsqueda
  - Podemos bajar los datos para análisis posteriores.
  - Veremos algunos ejemplos…

Ejemplos de consultas a *GenoMaiz*

La estructura de las tablas (nombre y variables que contienen) se puede ver AQUÍ.

En la parte de "Consulta Libre" es posible realizar consultas a la base de datos GenoMaiz utilizando el lenguaje "SQL". A

El formato general de una consulta es:

select <*variables*> from <*tabla*> where <*condicion(es)*>

las palabras "select", "from" y "where" son parte del lenguaje SQL (structured query language) que se utiliza para comun
de la tabla, mientras que <*condicion(es)*> son algunas restricciones para limitar el contenido que se presentará.

Las tablas que existen en la base de datos están listadas aquí. Para saber que <*variables*> existen en una tabla se debe de

Vamos a ir aprendiendo por ejemplos. El texto de cada ejemplo se presenta en rojo. Dicho texto se deberá copiar y pegar

Ejemplos que involucran contenido de una sola tabla

Tablas accesibles en *GenoMaiz*
Nombre (con liga) - *Descripción breve*
DNA_plates - Identificadores de placa y pozo por accesión y grupo
MAL - Marcadores y alelos encontrados en las accesiones
accession - Datos de accesion
fingerprint - Genotipo numérico (0 o 1) para cada accesión y grupo.
fingerprint_acc - Genotipo numérico (0 o 1) para cada accesión
lecturas - Lecturas originales para cada accesión, grupo y marcador
rareness - Coeficientes de "rareza" por accesión
distancias - Distancias genéticas y geográficas entre pares de accesiones
geodistance - Distancias geográficas entre pares de accesiones
geneticdistance - Distancias genéticas euclidianas entre pares de accesiones

Tabla DNA_plates

Actualmente con 2685 renglones.

Variable	Tipo	Descripción
DNAplate_id	int(5)	Identificador de placa
ESTADO	varchar(20)	Estado del país
DNA_plate	int(2)	Número de placa
DNA_well	varchar(3)	Identificador de pozo
accesion	varchar(10)	Accesión / grupo que ocupa el pozo

Regresar a tablas

Ejemplo:

select Estado, count(*) 'renglones', count(distinct accesion) 'accesiones' from DNA_plates group by estado

Tabla MAL (Marcador Alelo)

Actualmente con 274 renglones.

Variable	Tipo	Descripción
unid	int(5)	Identificador numérico
marcador	varchar(20)	Nombre del marcador
id_estado	varchar(2)	Identificador de estado
lim_inf	float	Límite inferior del alelo
lim_sup	float	Límite superior del alelo
alelo	int(2)	Alelo (número entero)

Regresar a tablas

Ejemplo:
select avg(lim_sup-lim_inf) 'ProRango', min(lim_sup-lim_inf) 'minRango', max(lim_sup-lim_inf) 'maxRango' from mal

Tabla accession

Actualmente con 967 renglones.

Variable	Tipo	Descripción
id	int(3)	Identificador numérico
acc_id	varchar(5)	Identificador de accesión (dos letras para estado, tres dígitos)
acc_id_ori	varchar(20)	Identificador original de la accesión (depende de fuente)
localidad	varchar(100)	Nombre de la localidad
tipo0	varchar(50)	Tipo secundario
tipo	varchar(50)	Tipo primario
tipo2	varchar(50)	Tipo terciario
color0	varchar(50)	Color secundario
color	varchar(50)	Color primario
lat_gra	float	Coordenadas (Latitud, grados)
lat_min	float	Coordenadas (Latitud, minutos)
lat_seg	float	Coordenadas (Latitud, segundos)
lon_gra	float	Coordenadas (Longitud, grados)
lon_min	float	Coordenadas (Longitud, minutos)
lon_seg	float	Coordenadas (Longitud, segundos)
asnm	int(4)	Altura sobre el nivel del mar en donde se colectó la accesión
UTM_X	float	Coordenadas UTM en X
UTM_Y	float	Coordenadas UTM en Y

Ejemplos de búsquedas en la tabla “accession”

- Cuantas accesiones
  - *“select count(*) from accession” o “select count(distinct acc_id) '#Acc' from accession”*
- Cuantas accesiones por localidad
  - *“select Localidad, count(*) '#Acc' from accession group by localidad order by count(*) desc”*
- Cuantas accesiones de cada color
  - *“select Color, count(*) '#Acc' from accession group by Color order by count(*) desc”*
- Y esta:
  - *“select Color, Localidad, count(*) '#Acc' from accession group by Color, Localidad order by count(*) desc”*

Ejemplos de búsquedas en la tabla “accession”

- ¿En que localidades hay maíz azul?
  - “select Localidad, count(*) '#A' from accession where color="Azul" or color0 = "azul" group by localidad order by count(*) desc”
- ¿Qué tipos de maíz fueron colectados a más de 3000 msnm y en que localidades?
  - “select Tipo, Localidad, round(avg(asnm)) 'PrASNN', count(*) '#A' from accession where asnm>3000 group by Tipo, Localidad order by count(*) desc”
- ¿y a no más de 100 msnm?
  - “select Tipo, Localidad, round(avg(asnm)) 'PrASNN', count(*) '#A' from accession where asnm<=100 and Localidad is not null group by Tipo, Localidad order by count(*) desc”
- Y esta:
  - “select distinctrow tipo0, tipo, tipo2 from accession where tipo like '%tuxp%' or tipo0 like '%tuxp%' or tipo2 like '%tuxp%'”

Tabla fingerprint

Actualmente con 1083 renglones.

Aparte de las variables "acc" (Accesión) y "batch" (Grupo)
esta tabla contiene datos numéricos <genotipo>= 0 o 1 para todas
las combinaciones encontradas de <Marcador>_<Alelo>;
por ejemplo la variable "PHI015_0" se refiere al alelo 0 (o nulo)
del marcador PHI015, etc.
Estos datos forman la huella genética del grupo de plantas
analizadas.

Field	Tipo	Descripción
acc	varchar(5)	Accesión
batch	int(1)	Grupo
PHI015_0	int(1)	<genotipo>
PHI015_63	int(1)	
PHI015_66	int(1)	

Tabla fingerprint_acc

Actualmente con 361 renglones.

Aparte de la variables "acc" (Accesión) esta tabla contiene
datos numéricos <genotipo>= 0, 1, 2 o 3 para todas
las combinaciones encontradas de <Marcador>_<Alelo>;
por ejemplo la variable "PHI015_0" se refiere al alelo 0 (o nulo)
del marcador PHI015, etc.
Estos datos forman la huella genética de la accesión y el genotipo
resulta de sumar los genotipos de los tres grupos analizados.

Variable	Tipo	Descripción
acc	varchar(5)	Accesión
PHI015_0	int(1)	<genotipo>
PHI015_63	int(1)	
PHI015_66	int(1)	
PHI015_69	int(1)	

Tabla lecturas

Actualmente con 58,474 renglones.

Variable	Tipo	Descripción
lec_id	int(5)	Identificador numérico
accesion	varchar(5)	Accesión ("acc" en otras tablas)
state_id	varchar(2)	Identificador de estado (dos letras)
acces_id	varchar(3)	Identificador numérico de accesión
id_group	int(1)	Identificador de grupo ("batch")
marker	varchar(15)	Nombre del marcador
alelo_bp	float	Alelo asignado
reading_bp	float	Lectura original

Regresar a tablas

Tabla rareness

Actualmente con 361 renglones.

Variable	Tipo	Descripción
accesion	varchar(5)	Accesión ("acc" en algunas tablas)
r_whole	double	Coeficiente de rareza en toda la colección
r_state	double	Coeficiente de rareza en el estado
r_maiz	double	Coeficiente de rareza en maíz solamente (no teocintle)
r_TE	double	Coeficiente de rareza en teocintle
r_PL	double	Coeficiente de rareza en Puebla
r_TB	double	Coeficiente de rareza en Tabasco
r_TL	double	Coeficiente de rareza en Tlaxcala
r_GT	double	Coeficiente de rareza en Guanajuato

Regresar a tablas

Recibidos (2.194) – octavio | (8114) LANGEBIO | RoundC | cinvestav-irapuato | Cons

computational.biology.langebio.cinvestav.mx/GenoMaiz/consulta.php

Aplicaciones | Bookmarks | DriveGmail | Google Académico | Save to Mendeley | LANGEBIO | Rou

de Datos

Consulta Base de Datos

Consulta: select * from accession, lecturas where accesion="PL001" and accesion=acc_id

Enviar

id	acc_id	acc_id_ori	localidad	tipo0	tipo	tipo2	color0	color	lat_gra	lat_min	lat_seg	lon_gra	lon_min
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54
689	PL001	34	Tetela de Ocampo	Arrocillo	Arrocillo	Arrocillo	Blanco	Blanco	19	50	37	97	54

Marcadores: PHI109188

Alelos: 0

Accessiones que presentan el alelo 0 del marcador PHI109188

Exportar

Regresar

Accession	Tipo	Color	Frec. PHI109188
GT017	Mushito	Blanco	1
GT032	Conico Norteno	Blanco	1
GT067	Conico Norteno	Negro	1
GT076			1
TL025	Chalqueno	Blanco	1
TL048	Chalqueno	Blanco	1

Genetic fingerprint

Exportar

Regresar

Accession	Tipo	Color	acc	PHI015_0	PHI015_63	PHI015_66	PHI015_69	PHI015_72	PHI015_75	PHI015_78	PHI015_80
TE004	Teosinte BALSAS		TE004	0	0	0	2	0	3	2	3
TE005	Teosinte BALSAS		TE005	0	1	0	1	0	3	3	1
TE006	Teosinte BALSAS		TE006	0	0	0	2	1	2	2	0
TE010	Teosinte BALSAS		TE010	0	3	0	3	2	1	1	2
TE015	Teosinte BALSAS		TE015	0	0	0	0	1	0	1	1
TE017	Teosinte BALSAS		TE017	0	1	0	1	0	0	1	2
TE018	Teosinte BALSAS		TE018	0	1	0	3	1	3	3	0
TE021	Teosinte BALSAS		TE021	0	0	0	0	0	1	0	3
TE023	Teosinte BALSAS		TE023	0	0	0	1	3	0	2	2

Tabla distancias

Actualmente con 64980 renglones.

La distancia genética euclidiana (dis_gen) entre pares de accesiones es la raiz cuadrada de las sumas de las diferencias al cuadrado entre huellas genéticas de las accesiones.

La distancia geográfica (dis_geo) se obtiene a partir de las coordenadas y está dada en Km.

Field	Tipo	Descripción
acc1	varchar(5)	Nombre de la primera accesión
tipo_ac1	varchar(50)	Tipo (tipo0) de la primera accesión
color_ac1	varchar(50)	Color (color) de la primera accesión
acc2	varchar(5)	Nombre de la segunda accesión
tipo_ac2	varchar(50)	Tipo (tipo0) de la segunda accesión
color_ac2	varchar(50)	Color (color) de la segunda accesión
dis_gen	double	Distancia Genética entre accesiones
dis_geo	double	Distancia Geográfica entre accesiones
prod_dist	double	producto de distancias

Colofón

- De momento, solo algunas de las tablas están a disposición del usuario
- La base de datos seguirá siendo incrementada y mejorada; para ello las críticas y sugerencias de los usuarios son muy valiosas.

Octavio Martínez de la Vega
omartine@langebio.cinvestav.mx