





Computational phenotyping of potential plant growth promoters (*Klebsiella*) isolates from INCAUCA fields

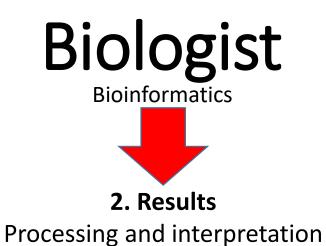
Luz Karime Medina Cordoba

02/08/18

Outline

1. Introduction

- Biologist vs Bioinformatics
- 2. General idea about microbes and the need of computational tools
- 3. What is Computational phenotyping?
 - Historical context of computational phenotyping
- 4. Computational phenotyping methods
 - Gene panels
 - > Blast
 - Microbial Identification and Characterization (MICRA)
 - > Traitar, the Microbial Trait Analyzer
 - Machine learning
- 5. Example of computational phenotyping (My project)
- 6. Computational Phenotyping methodology that I use for my project
- 7. Conclusion



of obtained results

3. Scientific articles

"Relevant" results are published in scientific journals



1. Experiments

Planning and carrying out

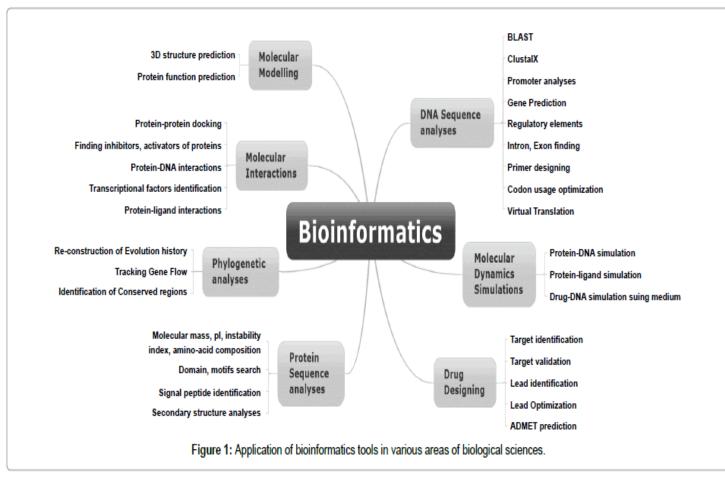
experiments (Lab work)





Applied bioinformatics

The application of computational techniques to understand and organize the information associated with biological macromolecules.



Microbes and the need of computational tools

- Bacteria are ubiquitous in our ecosystem and have a major impact on human health
- Diverse bacteria contribute with their unique capabilities to the functioning of such ecosystems
- Lab experiments to investigate those capabilities are laborintensive
- Computational tools help us to predicts traits of bacteria on the basis of their genomes



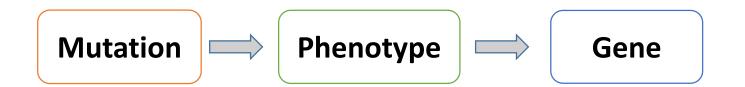


What is Computational phenotyping?

- Computational phenotyping is the use of software tools to describe the phenotypes of organisms using the genome sequencing
- ➢Good example of computational phenotyping is developing a software model to predicts minimum inhibitory concentrations for *Klebsiella pneumonie* antibiotics

Historical context

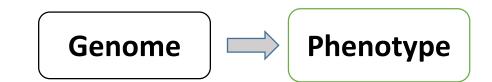
>(1900s) Forward genetics "Classic genetics": from phenotype to gene sequence



>(1970s) Reverse genetics "DNA sequencing era": From sequences to phenotype



>(2018) Reverse genomics "Next generation sequencing"



Computational phenotyping methods

- Gene panels
- ►Blast
- Microbial Identification and Characterization (MICRA)
- ➢ Traitar, the Microbial Trait Analyzer
- Machine learning

Gene panels

Contain a select set of genes or gene regions that have known or suspected associations with the phenotype under study

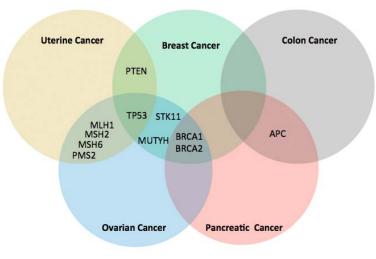
Advantages

- i. Facilitates the analysis of a group of genes of interest allowing identification of rare variants
- ii. Great approach when the database is not available
- iii. Easy to interpret results

Disadvantages

- i. Requires literature survey, which is time consuming
- ii. Some gene panels are not publically available

Н	uman Brea	st Cancer F	anel: 45 G	enes
ACVR1B	EP300	IRAK4	PBRM1	TP53
AKT1	ERBB2	ITCH	PCGF2	TRAF5
ATM	ERBB3	KMT2C	PIK3CA	WEE1
BAP1	ESR1	MAP2K4	PIK3R1	ZBED4
BRCA1	EXOC2	MAP3K1	PPM1L	ZNF226
BRCA2	EXT2	MDM2	PTEN	
CBFB	FBXO32	MUC16	PTGFR	
CDH1	FGFR1	MYC	RB1	
CDKN2A	FGFR2	NCOR1	RET	
EGFR	GATA3	NEK2	SEPT9	

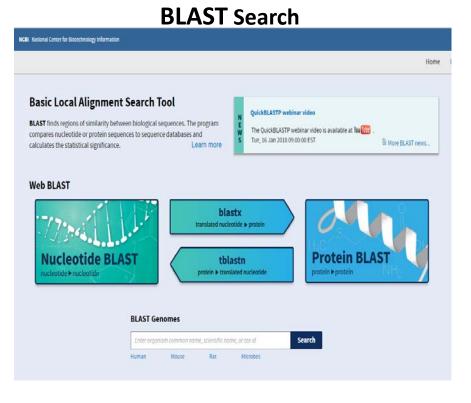


Blast (Basic Local Alignment search Tool)

- Blast tool is used to compare gene and protein sequences against other in public database
- It breaks the query and database sequences into fragments and seeks matches between them

Advantages

- i. Character string comparison against all the sequences on the target database
- ii. Rigorous statistics to identify statically significant matches
- iii. Helps to direct experimental design to prove the function



Blast (Basic Local Alignment search Tool)

- Advantages Find similar sequences in model organisms, which can be used to further study gene
- Compare complete genomes against each other to identify similarities and differences among organisms
- ii. Fast database searching

Disadvantages

- i. Requires some setup and computer expertise
- ii. Use GeneBank which is not well curated

Query (imput) sequence >Query1 >Query2 ACCCAAAAGCA **Results (output)** 50-80 80-200 40-50 <40 >=200 Querv 210 70 280 350 140

Microbial Identification and Characterization (MICRA)

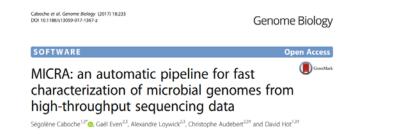
An automatic pipeline, available as a web interface, for microbial identification and characterization through reads analysis

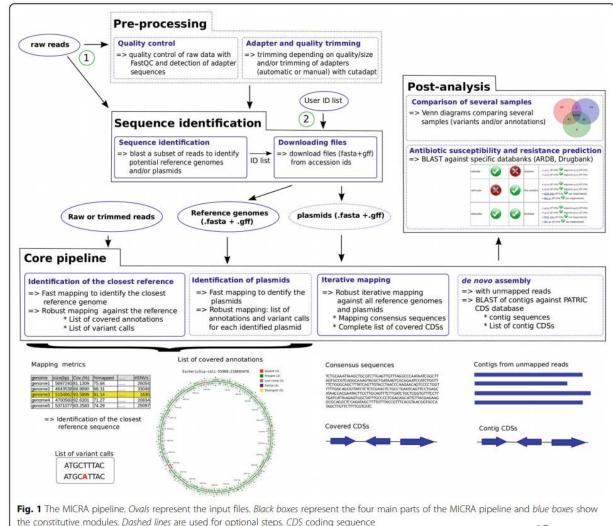
Advantages

- i. MICRA is freely available and user-friendly for both clinicians and biologists
- ii. Automatic analysis, requiring only reads as input.
- iii. MICRA offers the possibility of customizable analyses by giving access to a lot of setting parameters.
- iv. MICRA is fast (around 10 minutes in most cases)

Disadvantages

Lack of additional modules for a better interpretation of results





https://github.com/caboche/MICRA

Traitar, the Microbial Trait Analyzer

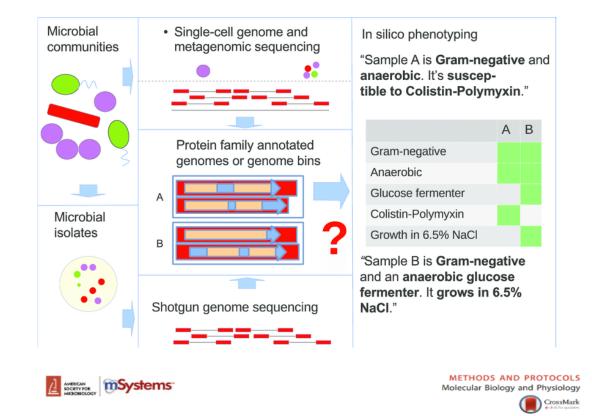
The microbial trait analyzer, which is a fully automated software package for deriving phenotypes from a genome sequence

Advantages

- i. Easy to use
- ii. Traitar provides phenotype classifiers to predict 67 traits morphology (antibiotic susceptibility, and enzymatic activities)
- iii. Can provide reliable insights into the metabolic capabilities of microbial community members even from partial genomes
- iv. It is freely available under the open-source

Disadvantages

i. The accuracy of the phenotype classification models



From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer

Aaron Weimann, ^a,b,c Kyra Mooren, ^a,c Jeremy Frank, ^d Phillip B. Pope, ^d Andreas Bremges, ^a,b Alice C. McHardy ^a,b,c

Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany^e; German Center for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Braunschweig, Germany^e; Department for Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany^e; Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway^d

Machine learning

Involves developing and deploying algorithms to provide a computer, a software program, or a process with the ability to learn without being explicitly programmed.

Advantages

- i. Supplementing data mining
- ii. Continuous improvements
- iii. Automation of tasks

Disadvantages

- i. Error diagnosis and correction
- ii. Problems with verification
- iii. Limitations of predictions

www.natore.com/scientificiepoits

SCIENTIFIC REPORTS

OPEN Developing an *in silico* minimum inhibitory concentration panel test for *Klebsiella pneumoniae*

Received: 27 September 2017 Marcus N Accepted: 12 December 2017 Olsen^{4,5}, J Published online: 11 January 2018 Yoo^{2,3} & J

Marcus Nguyen^{1,2,3}, Thomas Brettin^{2,3}, S. Wesley Long^{6,4,5}, James M. Musser^{4,5}, Randall J. Olsen^{3,5}, Robert Olson^{2,3}, Maulik Shukla^{2,3}, Rick L. Stevens^{2,2,4}, Fangfang Xia^{2,3}, Hyunseung Yoo^{2,3} & James J. Davis^{2,3}



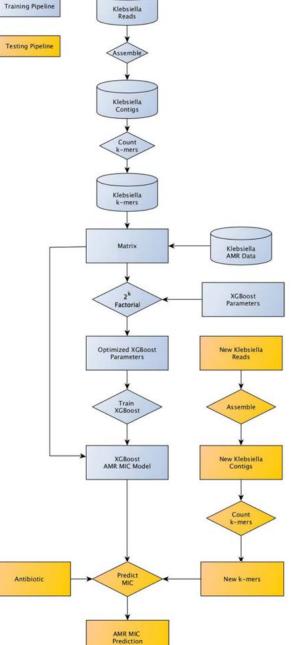


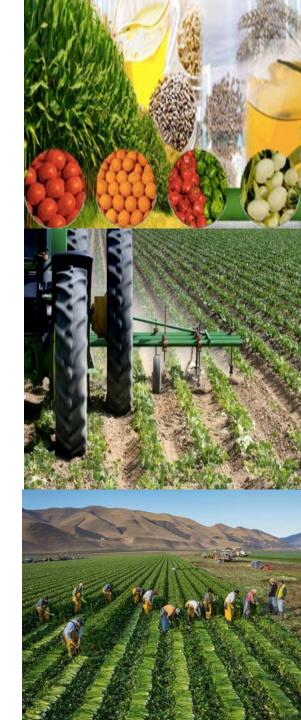
Figure 1. The pipeline used to optimize and train the XGBoost model using known data (blue), and to predict the MIC values for a new genome (yellow).

Example of Computational phenotyping

"Genomic characterization and prioritization of nitrogenfixing bacteria biofertilizers isolated from Colombian sugarcane fields"

Sustainable agriculture

- The increase in the world population and the environmental damage have brought as a consequence that more food is needed.
- > To feed the world population will be required that agricultural yields increase.
- > Demand of fertilizers, major cost for companies.
- Chemical fertilizers and biological fertilizers
- Biofertilizer that contains plant growth-promoting microorganisms
- Nitrogen-fixing bacteria or diazotrophs
 - plant growth-promoting microorganisms that fix nitrogen
- > Biological nitrogen fixation is a process carried out by nitrogen fixing bacteria.
 - Atmospheric dinitrogen (N2) is reduced into ammonia (NH3)
 - nitrogenase enzyme complex.





The research problem

- INCAUCA is a sugarcane company in Colombia, Colombia, South America, which plays a vital role in the economy of the country by supporting food, energy and fuel production.
- INCAUCA uses chemical fertilizers, such as urea, to promote sugarcane growth
- Chemical fertilizers may cause serious environmental problems
- To solve this problem, we propose a biological alternative to improve yields of crops using biofertilizer that contains plant growth-promoting microorganisms



Overall significance and goals of the study

- Previous studies have shown that sugarcane from INCAUCA fields harbors diverse plant growth promoting microorganisms (nitrogen-fixing bacteria), which have the potential to serve as biofertilizers.
- The success of biofertilizers depends on the capacity of the microorganism to adapt to the environmental conditions of the place where it is applied

> Endemic bacteria (natives of INCAUCA fields)

Characterizing endemic nitrogen-fixing bacteria from INCAUCA field, we will be able to know their potential as a biological fertilizer that promotes sugarcane growth in term of biomass

Field work

Objective : Isolate and characterize potential plant growth promoters (nitrogen fixing bacteria) from INCAUCA fields

- Sugar cane samples from INCAUCA fields were collected in May-June 2014
- Samples were taken from rhizosphere soil, roots, leaves & stem from different fields.
- Samples were transported to Georgia Tech for processing.



Wet lab work

Objetive: identify the culturable nitrogen fixing bacteria isolated from INCAUCA fields

- Pure cultures of nitrogen fixing bacteria from the sample were obtained (nitrogen free media)
- DNA was isolated from pure culture isolates.
- 16S rRNA and *nifH* amplification and sequencing was done these cultures.
- Diversity of bacterial species as determined from nitrogen fixation gene sequences
- *Klebsiella* is the second most abundant from metagenomic approach and the most abundant from the culture based approach
- We obtained 23 Isolates



Genomics & Bioinformatics

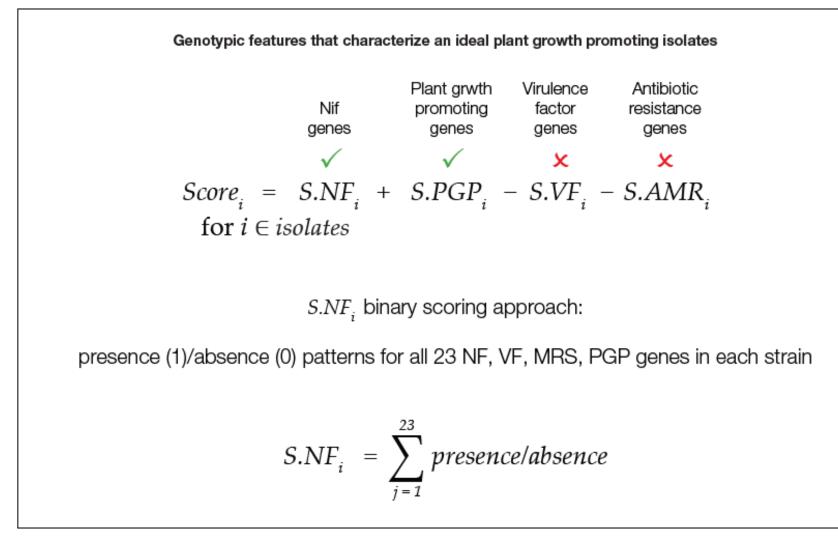
 Objective: Analyze whole genome sequence from 23 isolates in order to classify and prioritize potential plant growth promoting bacteria. We want strains that are predicted to have maximum benefit to the plants while presenting minimum risk to the environment, including local human populations.

Phenotypes of interest

 Define genotypic features that characterize an ideal plant growth promoting isolate

Genotypic features	Ideal plant growth promoting isolate
nif genes (nitrogen fixation)	\checkmark
Plant growth promotion genes	\checkmark
Virulence factor (VF) genes	X
Antibiotic resistance genes	X

Phenotypes of interest



Step 1. Literature survey – Creating gene panels

- i. Genes that have been implicated in these phenotypes
- ii. Collect gene sequences (From RefSeq / UniProt, anywhere)

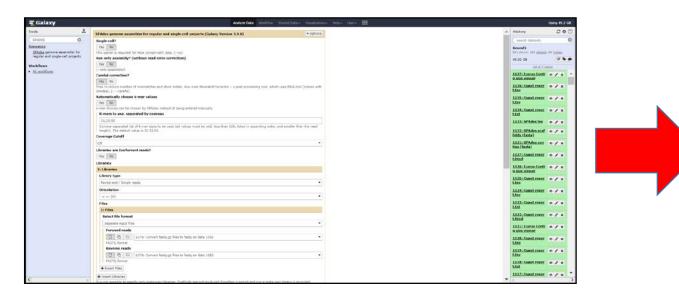
<i>nif</i> genes	Gene Symbol	Gene de	scription					
nifH	AB185_RS17065	Nitrogenase iron proteinrogenase iron protein	Nitrogenase iron proteinrogenase iron protein					
nifD	AB185_RS17060	Nitrogenase molybdenum-iron protein alpha cha	Nitrogenase molybdenum-iron protein alpha chain					
nifJ	BPR_RS01420	Structural- pyruvate:ferredoxin (flavodoxin) oxid	Structural- pyruvate:ferredoxin (flavodoxin) oxidoreductase					
nifF	AVCA6_RS00805	Flavodoxin, nifF						
nifA	blr2037	nif-specific regulatory protein						
nifL	AB185_RS16990	Nitrogen fixation negative regulator NifL						
nifE	AB185_RS17040	Nitrogenase iron-molybdenum cofactor biosynth	Nitrogenase iron-molybdenum cofactor biosynthesis protein NifE					
Plant growth pro	motion genes	Gene Name	Gene Symbol					
U		pqq	ASG52_RS18860					
		Glucose dehydrogenase gene homolog	YNL241C					
Phosphate solub		pstA	R2APBS1_RS07860					
•		pstB	KPHS_52970					
		pstC	AB185_RS07180					
		pstS	KPHS_53000					
AA production		ipdC	YE1222					
-		pvdO	PP_4215					
		pvdN	PP 4214					
		pvdP	PP_4212					

Step 2. Quality control

Galaxy	Analyze Data Workflow Shared Data+ Visualization+ Relp+		FastQC Report
1000000	FastQC Read Quality reports (Galaxy Version 0.67) Option		Mon 19 Jun 2017
ch tools	Short read data from your current history	search datasets	Convert_fastq.gz_files_to_fastq_on_data_47
ion of each het	① ② S76: Trimmomatic on K_S16_L001_R2_001.fastq.gz (R2 paired)	• Round2	
k trimfo trim FASTQ	Contaminant list	788 shown, 136 <u>deleted</u> 66 <u>hidden</u>	~
the Phred algorithm	C & Nothing selected	▼ 34,62 GB 🕑 🍽 🗩	Summary
<u>: hetv</u> regional ozvoosity	tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA	List of 2 poots	·
k mergefa merge two	Submodule and Limit specifing file	414: FastOC on data 3 ● / ▲ 54: RawData	
A/Q files	Q D Nothing selected a file that specifies which submodules are to be executed (default-all) and also specifies the thresholds for the each submodules warning	• 413: FastOC on data 3 • /	• Statistics
comp get the	a file that specifies which submodules are to be executed (default=ail) and also specifies the thresholds for the each submodules warning parameter	54: Webpage	
eotide composition of TA/Q	✓ Execute	412: FastQC on data 3 @	
mergepe interleave		53: RawData	 <u>Per base sequence quality</u>
unpaired FASTA/Q files	1 Purpose	411: FastOC on data 3 🗶 🖋	
paired-end file	FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipeline impression of whether your data has any problems of which you should be aware before doing any further analysis.	s. It provides a modular set of analyses which you can use to give a quick 53: Webpage	 <u>Per tile sequence quality</u>
k_fgchk fastq QC e/guality summary)	impression or whether your data has any problems or which you should be aware before doing any further analysis. The main functions of FastOC are:	410: FastOC on data 3 🔹 🖋	
dropse drop unpaired	Import of data from BAM, SAM or FastQ/FastQ.gz files (any variant),		• Per sequence guality scores
interleaved Paired End	 Providing a quick overview to tell you in which areas there may be problems 	409: FastQC on data 3 @ /	
TA/Q	 Summary graphs and tables to quickly assess your data Export of results to an HTML based permanent report 	408: FastOC on data 3 @ /	• • Per base sequence content
k_subseq extract equences from FASTA/Q	Offline operation to allow automated generation of reports without running the interactive application	51: RawData	
		407: FastOC on data 3 • /	• UPPer sequence GC content
mutfa point mutate	6 FastQC	51: Webpage	
at specified positions	This is a Galaxy wrapper. It merely exposes the external package FastOC which is documented at FastOC Kindly acknowledge it as well as this to		
seq common formation of FASTA/Q	processing.	<u>50: RawData</u>	• Per base N content
cutN cut sequence at	The contaminants file parameter was borrowed from the independently developed fastgowrapper contributed to the Galaxy Community Tool She	d by J. Johnson, Adaption to version 0.11.2 by T. McGowan. 405: FastOC on data 3 50: Webpage	
N	0		 Sequence Length Distribution
FASTQ reads by quality	Inputs and outputs Events In the best size to load for demonstration. We used at superson follows below for these is a traving burger	404: FastOC on data 3 🔹 🌶	
and length	EastOC is the best place to look for documentation - it's very good. A summary follows below for those in a tearing hurry.	contaminants information in the form of a tabudelimited file with 2 columns	 Sequence Duplication Levels
nomatic flexible read ning tool for Illumina	This wrapper will accept a Galaxy fastq, fastq.gz, sam or bam as the input read file to check. It will also take an optional file containing a list of name and sequence. As another option the tool takes a custom limits.bt file that allows setting the warning thresholds for the different module		
data	The tool produces a basic text and a HTML output file that contain all of the results, including the following:	402: FastQC on data 3 🔹 🖋	 Overrepresented sequences
C Read Quality reports	Basic Statistics	48: RawData	
ce comparison	Per base sequence quality Per sequence quality scores	401: FastOC on data 3 👁 🖋	• Adapter Content
0	Per base sequence content	48: Webpage	
20	Per base GC content Per sequence GC content	400: FastOC on data 3 🔹 🌶 47: RawData	Varia Cantant
<u>bls</u>	Per base N content		• <u>VKmer Content</u>
ows	Sequence Length Distribution Sequence Duplication Levels	399: FastQC on data 3 @ /	

Step 3. Assembly of the strains

Galaxy



Genome sequencing results for the 23 isolates

Sample ID	Genome Length	N50	L50	GC(%)	# of Contigs
SCK1	4,522,541	402,304	4	66.79	24
SCK2	5,231,439	417,927	5	59.33	53
SCK3	3,824,428	670,745	3	41.82	150
SCK4	4,511,030	223,239	8	66.79	55
SCK5	5,774,634	162,673	13	53.1	98
SCK6	6,094,823	117,689	15	56.73	294
SCK7	5,693,007	282,996	7	57.03	50
SCK8	5,695,902	281,292	9	57.03	50
SCK9	5,579,618	311,650	6	57.03	42
SCK10	5,591,472	614,324	3	57.03	34
SCK11	5,696,136	382,597	5	57.15	268
SCK12	5,817,089	176,655	10	57.02	79
SCK13	5,476,221	358,490	5	57.34	33
SCK14	5,465,811	300,899	5	57.34	41
SCK15	5,564,330	330,579	5	57.15	43
SCK16	5,795,921	478,592	3	54.06	84
SCK17	5,475,984	358,490	4	57.34	35
SCK18	5,476,135	422,400	3	57.34	32
SCK19	5,688,396	270,585	7	57.09	56
SCK20	5,500,801	82111	20	57.45	165
SCK21	5,324,920	112,078	15	55.26	100
SCK22	5,847,607	65,329	29	57.02	181
SCK23	5,919,817	400,012	7	57.01	270

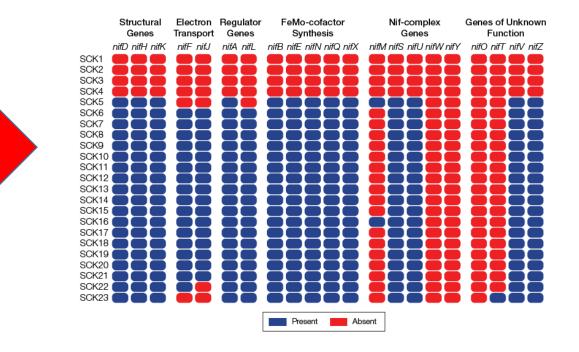
Functional Annotation of the strains

Step 4. Gene prediction and functional annotation

RAST (Rapid Annotation using Subsystem Technology)

nif genes involved in the fixation of atmospheric nitrogen

nað í Subsystem Technology anna 2	
The NMPOR, SEED based, prokenyted priorine anisotation service. For more information about The SEED prices vice https://www.com/seed.com/seed.com/seed.com/seed.com/seed.com/se	
Your Jobs	🍂 📢 Luz Hedina
NOTICE: PATRIC/RAST Tutorial Workshop, 2018-May 0709, 2017	
NTECHAST bases will be tartistical to tuburiul an 2016 May Co-Me, 24 Japanes Tables al Laboratory in the salabel of Chicage, IL: The vanishable will no for the active, will be tartistical and devote clearly barrispect working on on problems with his monthese. Texcology will be garantee assessing, assessing, assessing, and an instability and an instability and the salabel and the salabel and the salabel assessing, a	
voldado vill be linited to 35 popie se a finit-cannel finit-sanved basis. Tatevi va ne las for the venchade, pulti-tatevita en expensite fund the their own travel and togging regenerat. Viel for particities, van exercises (van Terresonal cold vice traces and cold van traces and cold vice traces and col	
orbino attentions must also separately III as the "Argener Value Report" from at <u>attenuing and any instruction as sen</u> as possible — expecting new U.S. Olizana, since they will need additional time to be disard. These List involves, church gluid gios as your "Sponter's email addivas." Please mate that your workshop registration will not be considered to be "Continued" and you have filled out the Valuer's respond form, and it has a diverged security.	
b To manifold RAST's load and view other news and statistics for RAST and the SEED, elease viet "The Daily SSED."	
NOT (Uspid Annotation using Sourceam Technology) is a MU-valurated service for annotabing complete or nearly complete batterial and and aead genomes. It provides high quality genome annotabions for these percents across the whole physic = Name Annotation University Sourceas constantions = Baultiferror Sourceas constantions = Baulti	penetic tree.
 Ibit Asis hand accretion interface (a per of myAs): Histor make interface (a per of myAs): Histor make interface (a per of myAs): Ibitor (a per of myAs):<	
(the number of more or less complete bacterial and archaeal perome sequences is constantly rising, the med for high quality actionated initial annotations in rising with it, in response to numerous requests for a SEED equality automated annot a community, mission bacteria and concerner estimated and the data international top quality general examples and the data annotation. RAS1 supports both the automated annot any makes the addated account of estimations of high quality general examples and the data annotation of high quality general examples an data annotation of high quality general examples and the data annotation of high quality general examples and the data annotation of high quality general examples and the data annotation of high quality general examples and the data a	
osse note that while the SEED environment and SEED data structures (most prominently <u>Flotters</u>) are used to compute the automatic annotations, the data is NOT added into the SEED automatically. Users can however request inclusion e mpleted, generate can be downloaded in a variety of formate or viewed online. The genome annotation provided dees include a mapping of genes to <u>subsystems</u> and a metabolic reconstruction.	f a their genome in the SEED. Once annotation is
be able to contact you once the computation is finished and in case user intervention is required, we request that users register with small address.	
you use the results of this annotation in your work, please cite:	
The 600 E Survey: Stand Analysis of Englished Employee Standbards Edwordspin Survey (Cases 6), Solid Cases 6, Solid Cases 6, March 6, Salesh 6, Salesh 6, Salesh 6, Salesh 7, Sa	ssieva O, Vonstein V, Wilke A, Zagnitico O.



BLAST against my gene panels Step 5. Finding genes of interest using BLAST

roject View 🗉 🗕 🗆 🗙	Carach Min				es [megablast] X										
0.2.11.1	/	1			ric Table View) [New Project]										
. New Project (*)		Sequence alginite													
Exact Match V Bearch:															
BLAST Results		Query	Subject	Query Start	Query Stop Query Stran	d Subject Start Su	ubject Stop Subject Strand	Identity	Coverage	Mismatches	Gaps	Gap Bases	Score	E-Value	Query Defline Jubject Deflin
Mixed sequence alignments, 22 entries [megablast]		-	,	-	1077 +			,	96.471681			16	594		
fig:66666666.275578.peg.95 (5364 components)		fig:6666666.27! E			1382 +	727	1041 +	85.687204 71.969697	45.197740	135 165	9	20	594 95	0.000000	v
Data Loaders		fig:6666666.27! E			1382 +	727	1376 +	71.969697	45.197740	165	17		95	0.000000	
DB: VFDB_setA_nt.fasta		fig:6666666.27! E			2583 +	133	2604 +	73.245086	45.197740 93.807339	628	34		472	0.000000	
All Views		fig:6666666.27! E			711 +	46	756 +	84.528833	100.000000	110	54	0	381	0.000000	
T A: Mixed sequence alignments, 22 entries [megablast] (Generic	5	fig:6666666.27! E		1	1644 +	40	1644 +	88.578372	99.878345	110	4	0	1080	0.000000	
Data Sources	7	fig:6666666.27! E			1644 +	2	77 +	96.000000	4.562044	104	4	4	66	0.000000	
BAM	9	fig:6666666.27! E			2526 +	1	2526 +	87.663108	99.881235	306	6	6	1590	0.000000	
	9	fig:6666666.27! E			669 +	1	669 +	85.522388	99.850523	95	2	2	378	0.000000	
		fig:6666666.27! E			584 +	1	584 +	89.914530	99.149660	57	2		407	0.000000	
NCBI Net BLAST	11	fig:6666666.27! E			543 +	49	591 +	87.661142	100.000000	67	0	0	342	0.000000	
	12	fig:6666666.27! E			392 +	254	419 +	76.785714	15.937804	35	3	4	49	0.000000	
	13	fig:6666666.27! E			756 +	22	747 +	81.069959	95.634921	132	6		312	0.000000	
	14	fig:6666666.27! E			852 +	1	858 +	81.976744	99.765258	145	7	-	390	0.000000	
	15	fig:6666666.27! E		4	1589 +	7	1592 +	78.477854	97.574627	311	27		551	0.000000	
	16	fig:6666666.27! E		968	1350 +	995	1377 +	72.820513	23.383085	92	13		65	0.000000	
	17	fig:6666666.27! E			1026 +	1045	1092 +	93.750000	2.985075	3	0	0	39	0.000000	
		fig:6666666.27! E			1176 +	13	1188 +	74.285714	98.809524	278	25	28	258	0.000000	
		fig:6666666.27! E			993 +	112	993 +	79.458239	88.418933	174	7		336	0.000000	
		fig:6666666.27! E			782 +	17	782 +	81.168831	95.849057	137	8	8	331	0.000000	
		fig:6666666.27! E			3719 +	301	3717 +	75.108288	86.888202	772	80	90	832	0.000000	
		fig:6666666.27! E			2229 +	96	2241 +	79.390018	94.257515	383	41	63	794	0.000000	

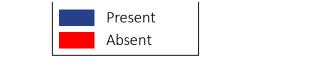
Step 6. Interpreting my results

> What makes a gene "present" in the genome?

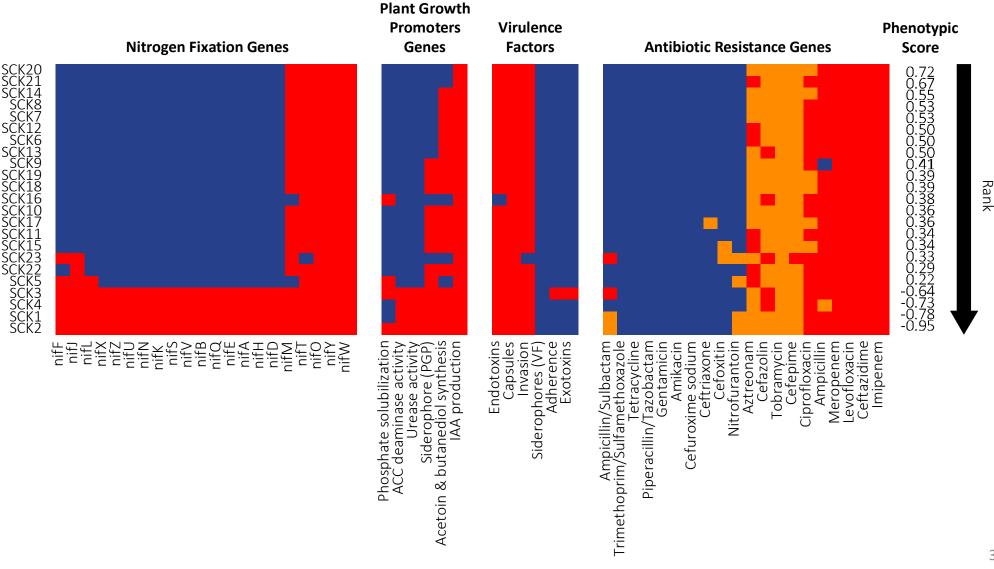
Identity	Coverage	Gaps	Score	E-Value	Genes
81.7204	8.61111	0	42	7.27E-16	gb AJ011502 Klebsiella pneumoniae OmpK37 Klebsiella pneumoniae
99.6516	100	0	852	0	gb AM850914 Klebsiella pneumoniae
99.5354	100	0	849	0	gb AM850909 Klebsiella pneumoniae
99.5354	100	0	849	0	gb AY743416 Klebsiella pneumoniae
99.4193	100	0	846	0	gb AM850912 Klebsiella pneumoniae
99.4193	100	0	846	0	gb AY037780 Klebsiella pneumoniae
94.2149	100	0	898	0	gb AJ318073.1 Klebsiella pneumoniae acrA Klebsiella pneumoniae
77.2586	29.3447	7	96	6.79E-46	gb AJ011502 Klebsiella pneumoniae OmpK37 Klebsiella pneumoniae
97.2603	56.5891	0	67	9.76E-31	gb AJ011502 Klebsiella pneumoniae OmpK37 Klebsiella pneumoniae
95.7333	100	0	981	0	gb/AJ011502/Klebsiella pneumoniae OmpK37 Klebsiella pneumoniae

> Empirical cutoffs (e.g. \geq 75% identity over \geq 75% of the length)

>What is the minimum set of genes needed for the phenotype







Conclusions

- Computational phenotyping software helps predict the phenotypes of organisms using only their genome sequences
- Computational phenotyping tools are more useful if they scale from few to many genomes
- Computational phenotyping can guide wetlab research by highlighting traits of interest, reducing the amount of wet lab work required