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

Biologist

Bioinformatics

1. Experiments

Planning and carrying out experiments (Lab work)

2. Results

Processing and interpretation of obtained results

3. Scientific articles

“Relevant” results are published in scientific journals

Applied bioinformatics

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

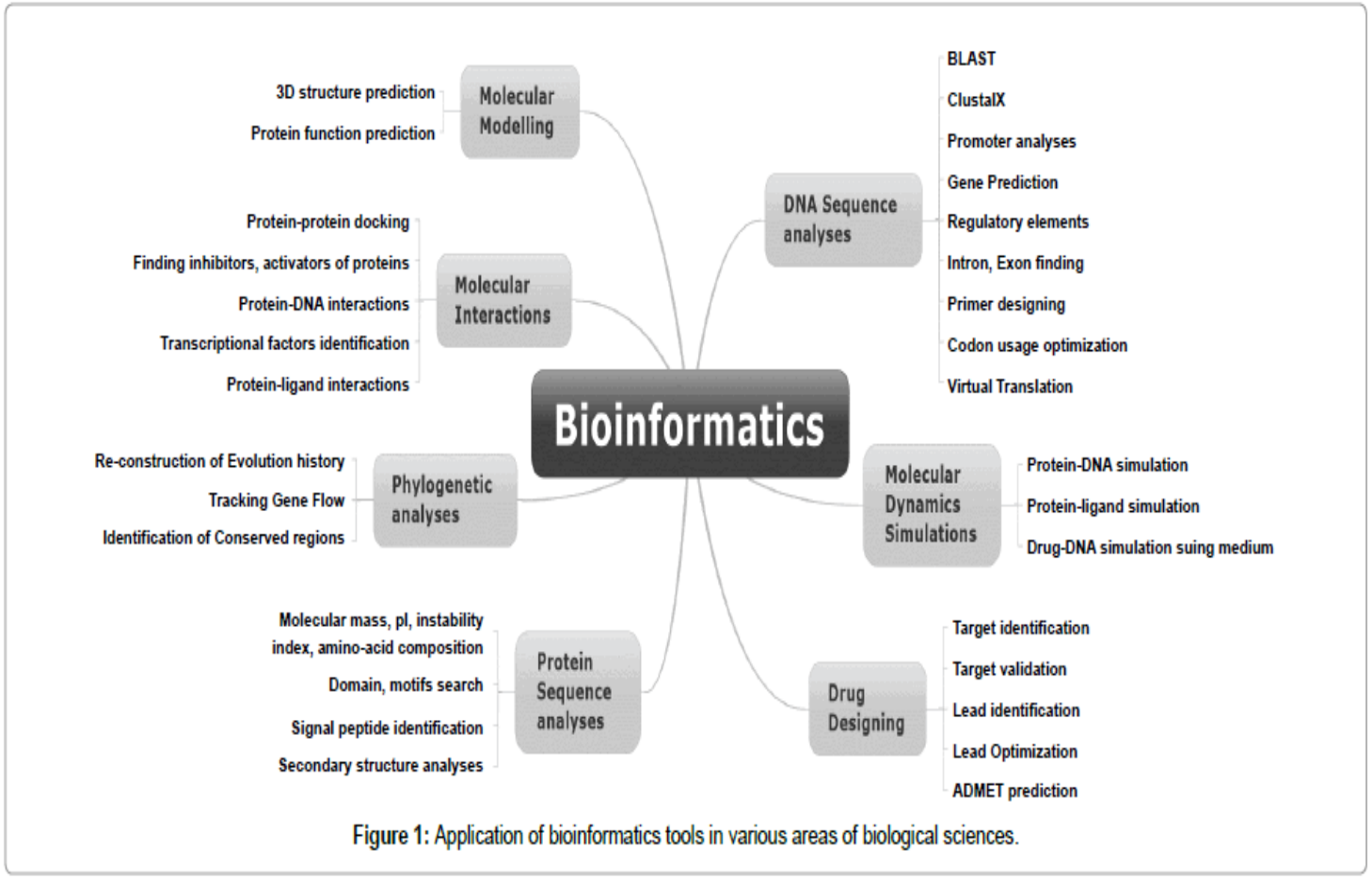


Figure 1: Application of bioinformatics tools in various areas of biological sciences.

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 labor-intensive
- 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

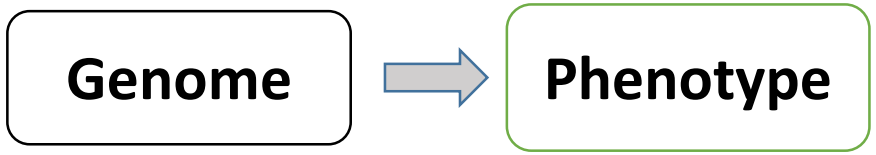
Historical context

Timeline

- **(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

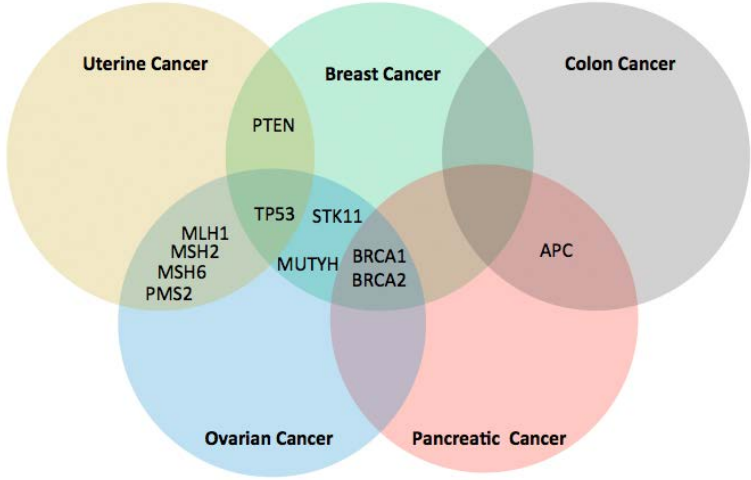
- Contain a select set of genes or gene regions that have known or suspected associations with the phenotype under study
- **Advantages**

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

Human Breast Cancer Panel: 45 Genes				
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	

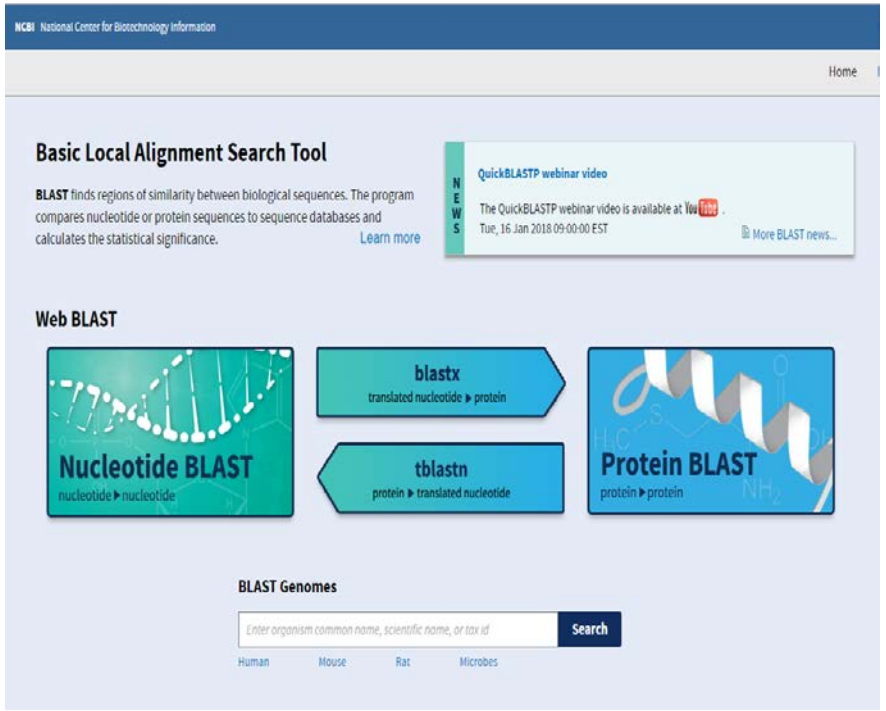


Blast (Basic Local Alignment search Tool)

- 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 Search



Blast (Basic Local Alignment search Tool)

- **Advantages** Find similar sequences in model organisms, which can be used to further study gene

i. Compare complete genomes against each other to identify similarities and differences among organisms

ii. Fast database searching

- **Disadvantages**

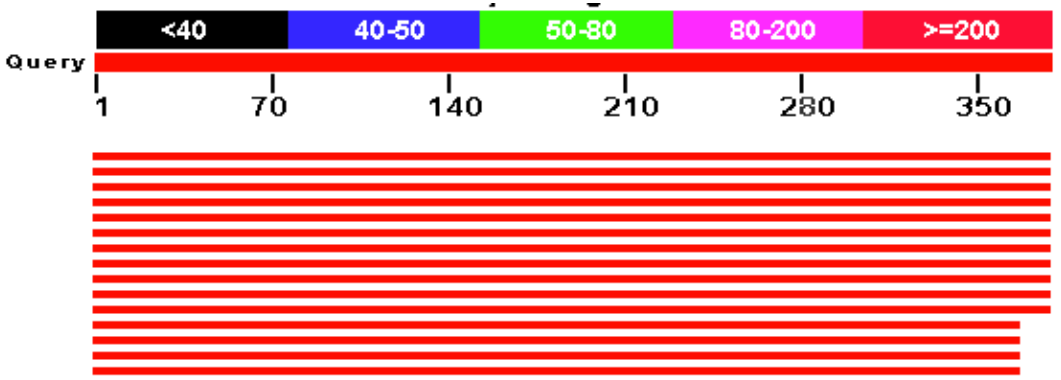
i. Requires some setup and computer expertise

ii. Use GeneBank which is not well curated

Query (imput) sequence

```
>Query1
ATGTCCTTTTTCTGTTCTATCTCCGGCGAACCTCCTCAGGACCCTGTCGTGT
        CCTCGAAGTCCGGCCATGTGTATGAGAGGAGGCTCATCCTGAAGTACATCACCGAGAATGGGACGGACCC
        AATTACTGGGGATAAGTTGGAGGAGGGAGACCTTATAACCGTGAAAGCGA
>Query2
ACCCAAAAGCAG
        CGCCGCCGCGGCCGCCAAATGCGAGCTCCATCCCTGCGCTCCTCCAAACCCTCCAAAATGAATGGGACGC
        GGCTATGCTCGAGACTTTTTCGCTCAAGCAGCAATATAACTCCCTCCGTCAGGAGCTCAGTTATGCCCTC
        TACGCACAAGACGCGGCGCAACGCGTCGTGGCGCGTCTCATCAAGGAGCGTGACGCAGCACGAGA
```

Results (output)



Microbial Identification and Characterization (MICRA)

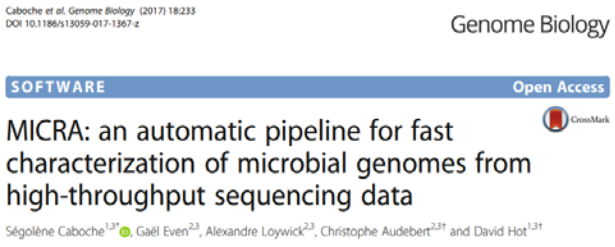
- An automatic pipeline, available as a web interface, for microbial identification and characterization through reads analysis

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


Caboche et al. Genome Biology (2017) 18:233
DOI 10.1186/s13059-017-1367-z

Genome Biology

SOFTWARE Open Access

MICRA: an automatic pipeline for fast characterization of microbial genomes from high-throughput sequencing data

Ségolène Caboche[1,3*], Gaël Even[2,3], Alexandre Loywick[2,3], Christophe Audebert[2,3†] and David Hot[1,3†]

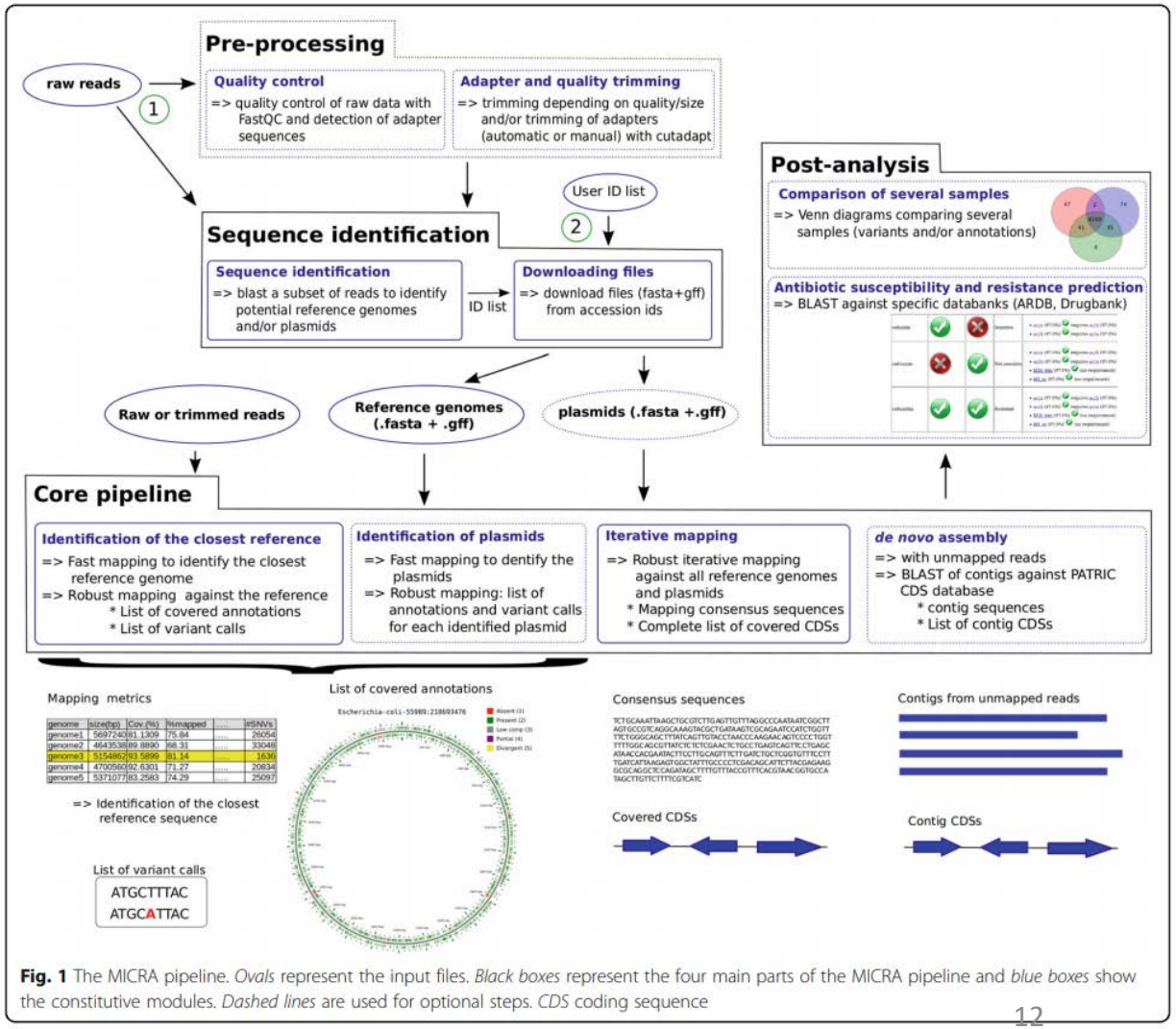


Fig. 1 The MICRA pipeline. *Ovals* represent the input files. *Black boxes* represent the four main parts of the MICRA pipeline and *blue boxes* show the constitutive modules. *Dashed lines* are used for optional steps. *CDS* coding sequence

https://github.com/caboche/MICRA

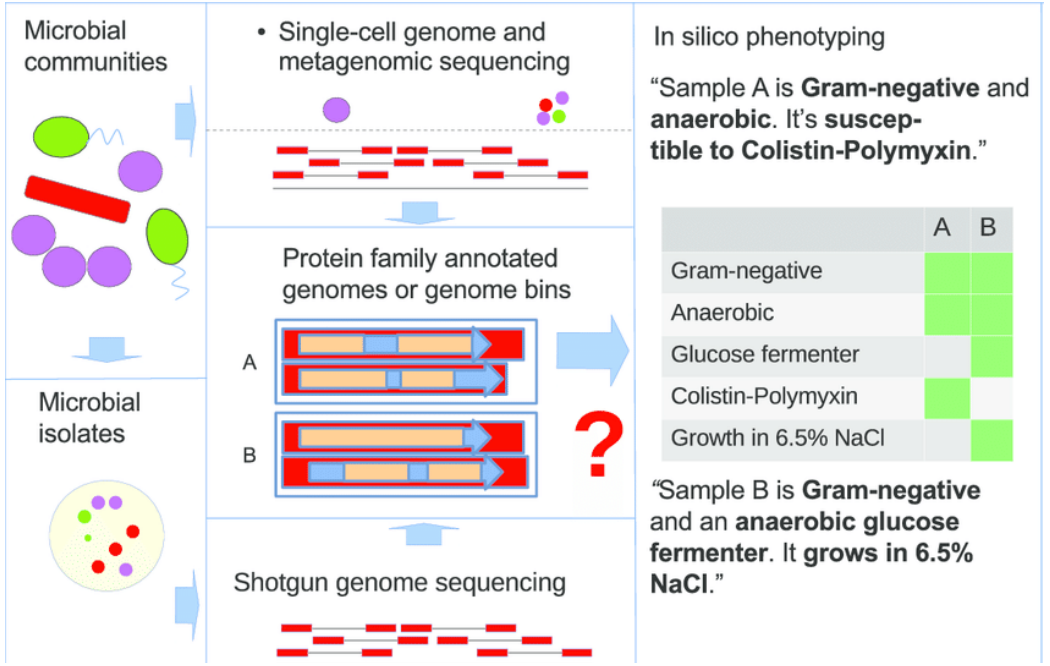
Traitar, the Microbial Trait Analyzer

- 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



mSystems
METHODS AND PROTOCOLS
Molecular Biology and Physiology

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[a]; German Center for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Braunschweig, Germany[b]; Department for Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany[c]; 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

Machine learning

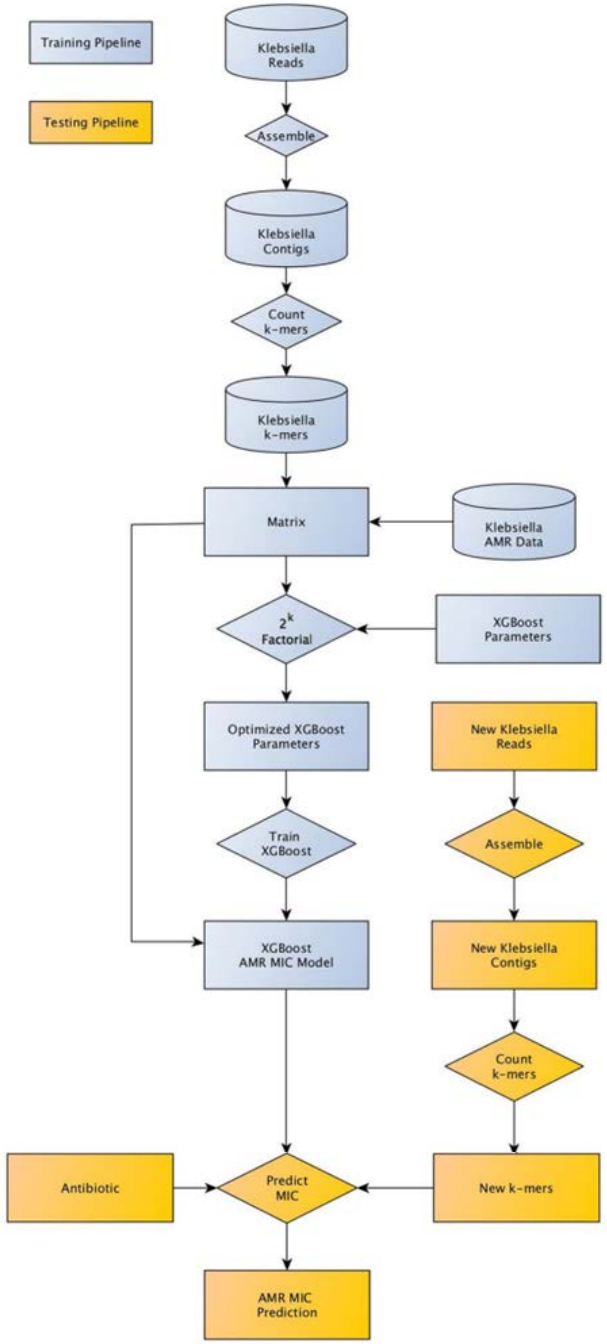


Figure 1. The pipeline used to optimize and train the XGBoost model using known data (blue), and to predict the MIC values for a new genome (yellow).

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

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

Received: 27 September 2017
Accepted: 12 December 2017
Published online: 11 January 2018

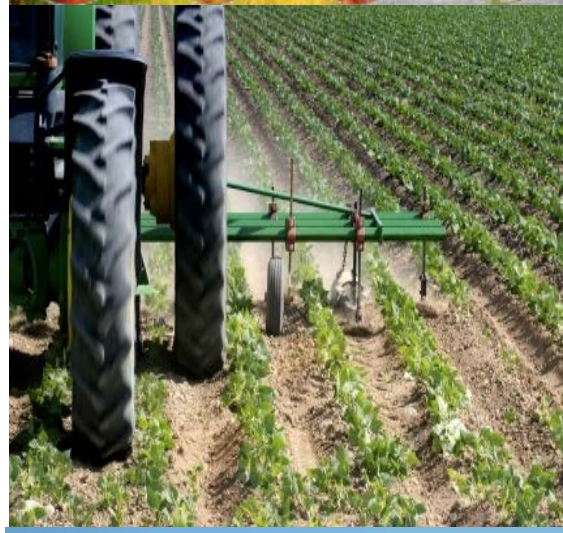
Marcus Nguyen[1,2,3], Thomas Brettin[2,3], S. Wesley Long[4,5], James M. Musser[4,5], Randall J. Olsen[4,5], Robert Olson[2,3], Maulik Shukla[2,3], Rick L. Stevens[2,3,6], Fangfang Xia[2,3], Hyunseung Yoo[2,3] & James J. Davis[2,3]

Example of Computational phenotyping

"Genomic characterization and prioritization of nitrogen-fixing 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 (N_2) is reduced into ammonia (NH_3)
  - 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

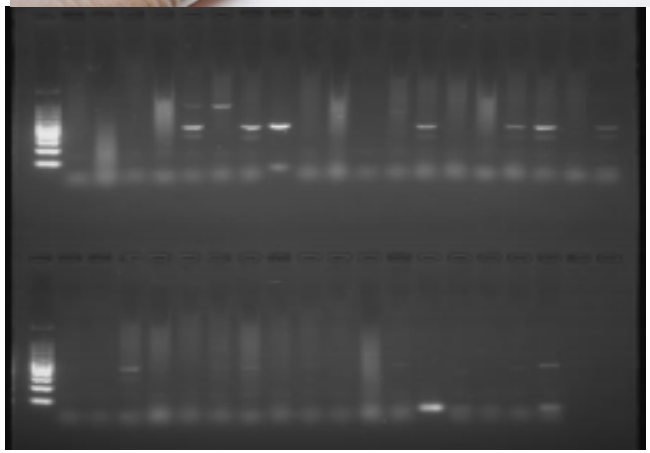
Objective : Isolate and characterize potential plant growth promoters (nitrogen fixing bacteria) from INCAUCA fields

- 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)	✓
Plant growth promotion genes	✓
Virulence factor (VF) genes	X
Antibiotic resistance genes	X

Phenotypes of interest

Genotypic features that characterize an ideal plant growth promoting isolates

Nif genes	Plant grwth promoting genes	Virulence factor genes	Antibiotic resistance genes
✓	✓	✗	✗

$$Score_i = S.NF_i + S.PGP_i - S.VF_i - S.AMR_i$$
$$\text{for } i \in isolates$$

$S.NF_i$ binary scoring approach:

presence (1)/absence (0) patterns for all 23 NF, VF, MRS, PGP genes in each strain

$$S.NF_i = \sum_{j=1}^{23} presence/absence$$

Computational Phenotyping methodology

Step 1. Literature survey – Creating gene panels

i. Genes that have been implicated in these phenotypes

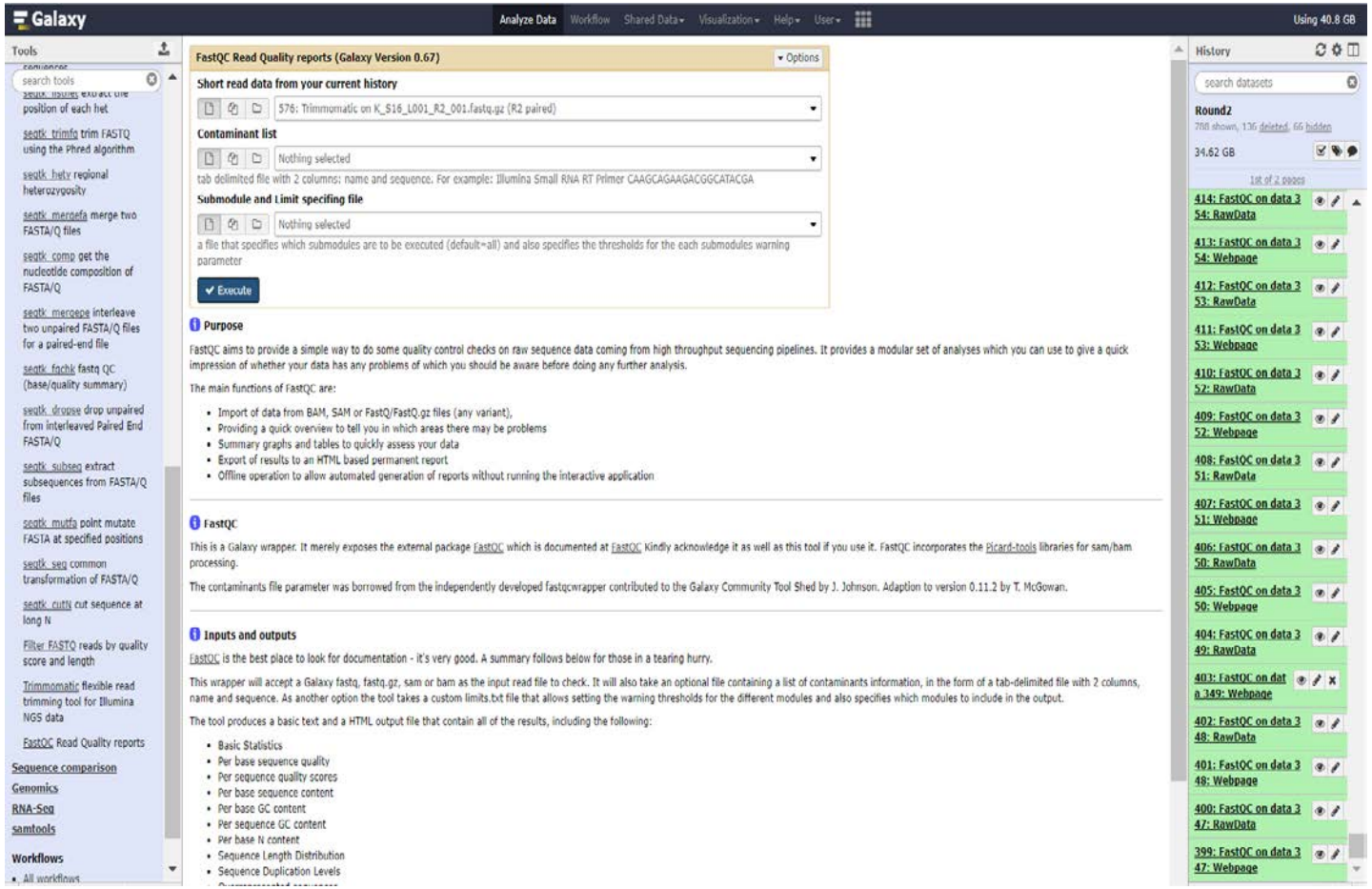
ii. Collect gene sequences (From RefSeq / UniProt, anywhere)

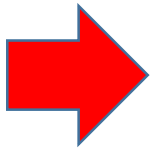
***nif* genes**	**Gene Symbol**	**Gene description**
nifH	AB185_RS17065	Nitrogenase iron proteinrogenase iron protein
nifD	AB185_RS17060	Nitrogenase molybdenum-iron protein alpha chain
nifJ	BPR_RS01420	Structural- pyruvate:ferredoxin (flavodoxin) oxidoreductase
nifF	AVCA6_RS00805	Flavodoxin, nifF
nifA	blr2037	nif-specific regulatory protein
nifL	AB185_RS16990	Nitrogen fixation negative regulator NifL
nifE	AB185_RS17040	Nitrogenase iron-molybdenum cofactor biosynthesis protein NifE

Plant growth promotion genes	**Gene Name**	**Gene Symbol**
Phosphate solubilization	*pqq*	ASG52_RS18860
	Glucose dehydrogenase gene homolog	YNL241C
	pstA	R2APBS1_RS07860
	pstB	KPHS_52970
	pstC	AB185_RS07180
	pstS	KPHS_53000
IAA production	*ipdC*	YE1222
	pvdO	PP_4215
	pvdN	PP_4214
	pvdP	PP_4212

Computational Phenotyping methodology

Step 2. Quality control

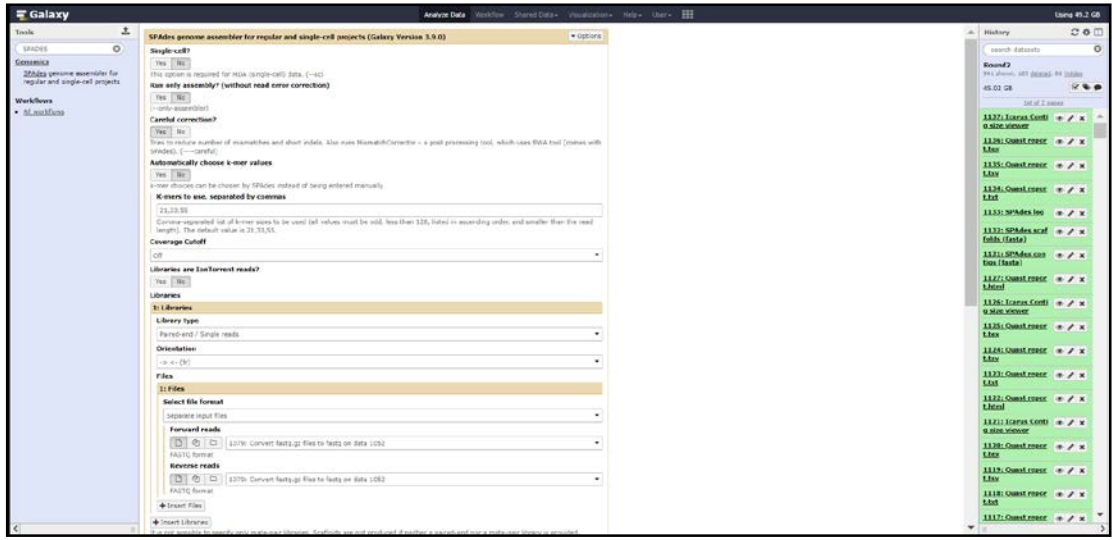




Computational Phenotyping methodology

Step 3. Assembly of the strains

Galaxy



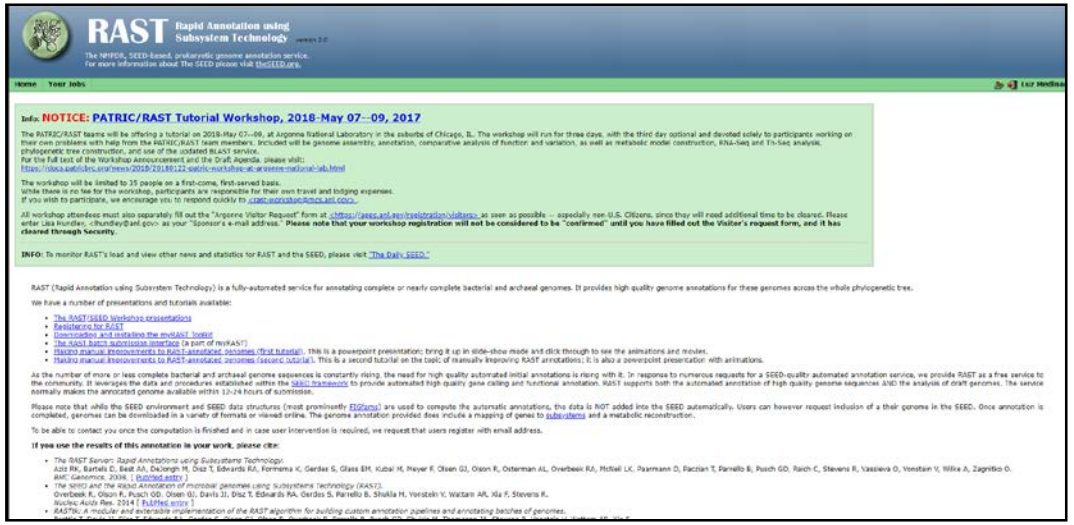
Genome sequencing results for the 23 isolates

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	40

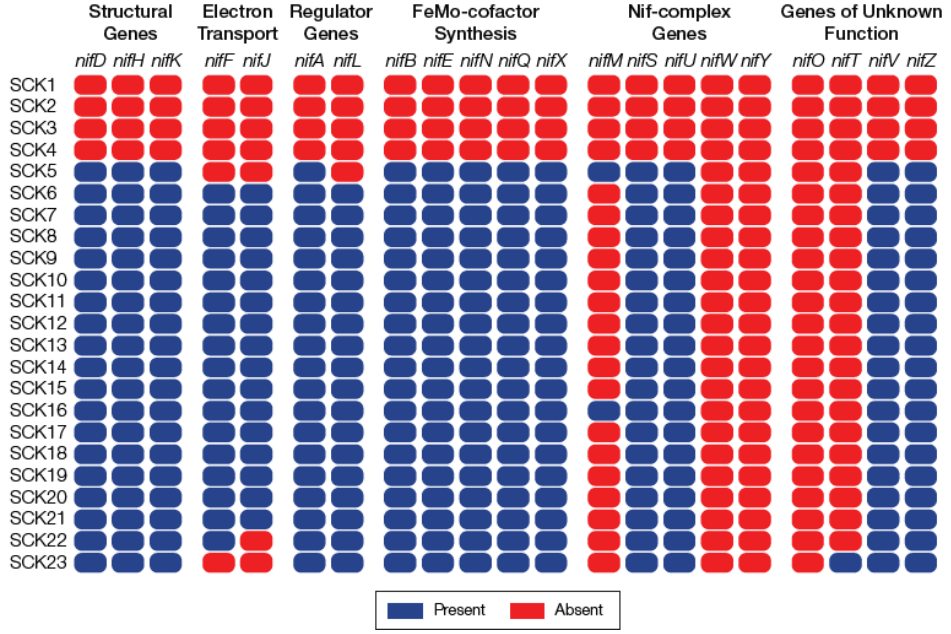
Functional Annotation of the strains

Step 4. Gene prediction and functional annotation

RAST (Rapid Annotation using Subsystem Technology)

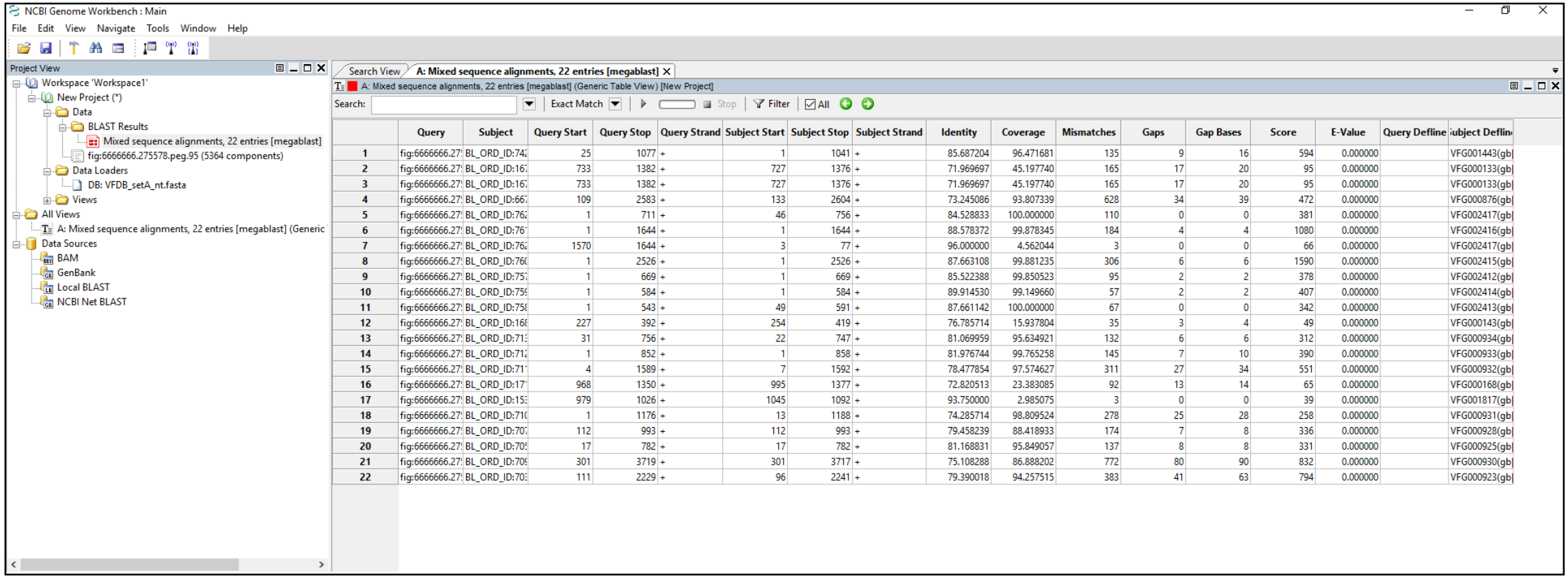


nif genes involved in the fixation of atmospheric nitrogen



BLAST against my gene panels

Step 5. Finding genes of interest using BLAST



	Query	Subject	Query Start	Query Stop	Query Strand	Subject Start	Subject Stop	Subject Strand	Identity	Coverage	Mismatches	Gaps	Gap Bases	Score	E-Value	Query Defline	ubject Defli
1	fig:6666666.27	BL_ORD_ID:74	25	1077	+	1	1041	+	85.687204	96.471681	135	9	16	594	0.000000		VFG001443(gb|
2	fig:6666666.27	BL_ORD_ID:16	733	1382	+	727	1376	+	71.969697	45.197740	165	17	20	95	0.000000		VFG000133(gb|
3	fig:6666666.27	BL_ORD_ID:16	733	1382	+	727	1376	+	71.969697	45.197740	165	17	20	95	0.000000		VFG000133(gb|
4	fig:6666666.27	BL_ORD_ID:66	109	2583	+	133	2604	+	73.245086	93.807339	628	34	39	472	0.000000		VFG000876(gb|
5	fig:6666666.27	BL_ORD_ID:76	1	711	+	46	756	+	84.528833	100.000000	110	0	0	381	0.000000		VFG002417(gb|
6	fig:6666666.27	BL_ORD_ID:76	1	1644	+	1	1644	+	88.578372	99.878345	184	4	4	1080	0.000000		VFG002416(gb|
7	fig:6666666.27	BL_ORD_ID:76	1570	1644	+	3	77	+	96.000000	4.562044	3	0	0	66	0.000000		VFG002417(gb|
8	fig:6666666.27	BL_ORD_ID:76	1	2526	+	1	2526	+	87.663108	99.881235	306	6	6	1590	0.000000		VFG002415(gb|
9	fig:6666666.27	BL_ORD_ID:75	1	669	+	1	669	+	85.522388	99.850523	95	2	2	378	0.000000		VFG002412(gb|
10	fig:6666666.27	BL_ORD_ID:75	1	584	+	1	584	+	89.914530	99.149660	57	2	2	407	0.000000		VFG002414(gb|
11	fig:6666666.27	BL_ORD_ID:75	1	543	+	49	591	+	87.661142	100.000000	67	0	0	342	0.000000		VFG002413(gb|
12	fig:6666666.27	BL_ORD_ID:16	227	392	+	254	419	+	76.785714	15.937804	35	3	4	49	0.000000		VFG000143(gb|
13	fig:6666666.27	BL_ORD_ID:71	31	756	+	22	747	+	81.069959	95.634921	132	6	6	312	0.000000		VFG000934(gb|
14	fig:6666666.27	BL_ORD_ID:71	1	852	+	1	858	+	81.976744	99.765258	145	7	10	390	0.000000		VFG000933(gb|
15	fig:6666666.27	BL_ORD_ID:71	4	1589	+	7	1592	+	78.477854	97.574627	311	27	34	551	0.000000		VFG000932(gb|
16	fig:6666666.27	BL_ORD_ID:17	968	1350	+	995	1377	+	72.820513	23.383085	92	13	14	65	0.000000		VFG000168(gb|
17	fig:6666666.27	BL_ORD_ID:15	979	1026	+	1045	1092	+	93.750000	2.985075	3	0	0	39	0.000000		VFG001817(gb|
18	fig:6666666.27	BL_ORD_ID:71	1	1176	+	13	1188	+	74.285714	98.809524	278	25	28	258	0.000000		VFG000931(gb|
19	fig:6666666.27	BL_ORD_ID:70	112	993	+	112	993	+	79.458239	88.418933	174	7	8	336	0.000000		VFG000928(gb|
20	fig:6666666.27	BL_ORD_ID:70	17	782	+	17	782	+	81.168831	95.849057	137	8	8	331	0.000000		VFG000925(gb|
21	fig:6666666.27	BL_ORD_ID:70	301	3719	+	301	3717	+	75.108288	86.888202	772	80	90	832	0.000000		VFG000930(gb|
22	fig:6666666.27	BL_ORD_ID:70	111	2229	+	96	2241	+	79.390018	94.257515	383	41	63	794	0.000000		VFG000923(gb|

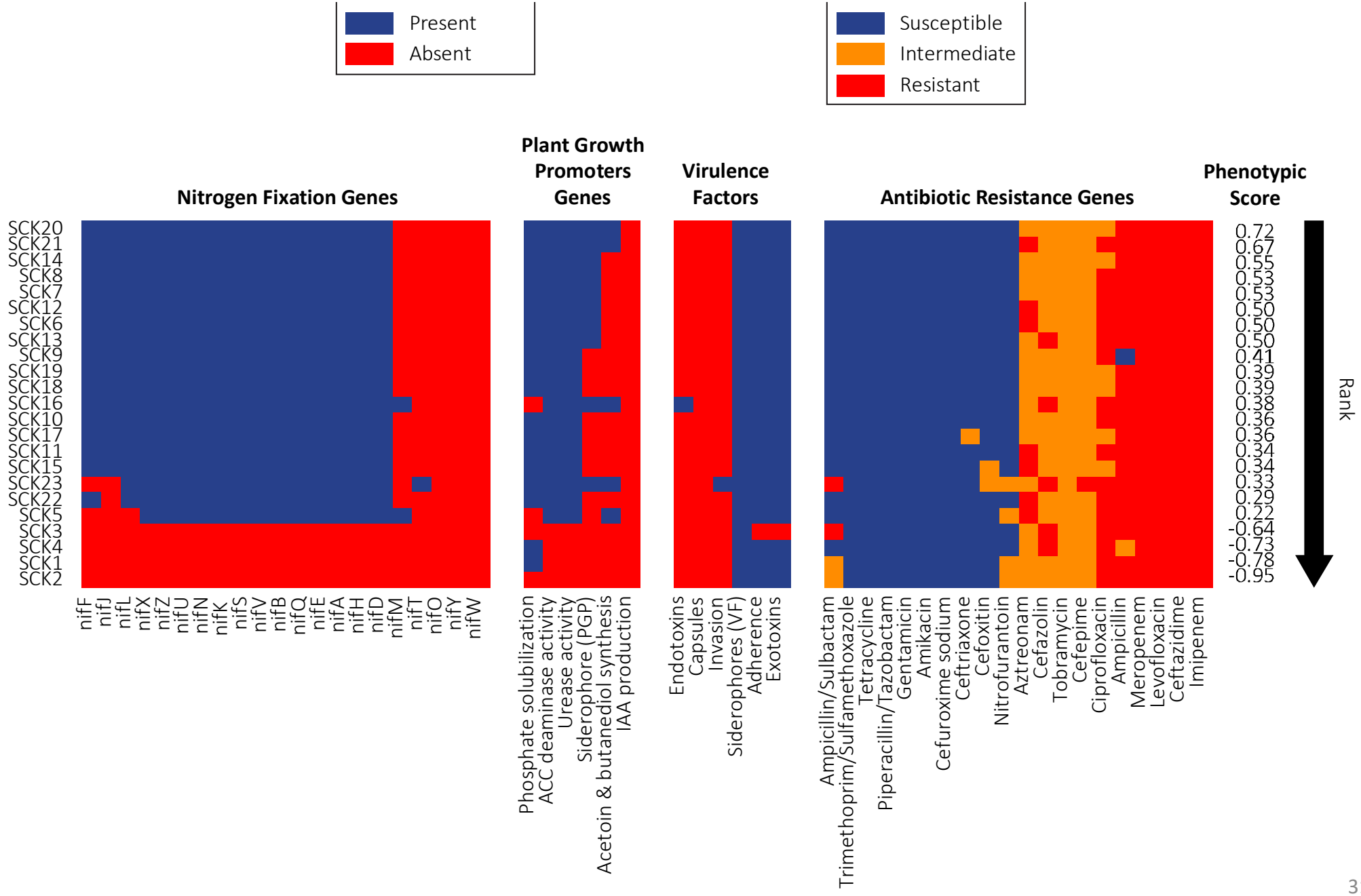
Computational Phenotyping methodology

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. ≥ 75% identity over ≥ 75% of the length)
- What is the minimum set of genes needed for the phenotype


Present
Absent
Susceptible
Intermediate
Resistant
Nitrogen Fixation Genes
Plant Growth Promoters Genes
Virulence Factors
Antibiotic Resistance Genes
Phenotypic Score
SCK20
SCK21
SCK14
SCK8
SCK7
SCK12
SCK6
SCK13
SCK9
SCK19
SCK18
SCK16
SCK10
SCK17
SCK11
SCK15
SCK23
SCK22
SCK5
SCK3
SCK4
SCK1
SCK2
0.72
0.67
0.55
0.53
0.53
0.50
0.50
0.50
0.41
0.39
0.39
0.38
0.36
0.36
0.34
0.34
0.33
0.29
0.22
-0.64
-0.73
-0.78
-0.95
Rank
nifF
nifJ
nifL
nifX
nifZ
nifU
nifN
nifK
nifS
nifV
nifB
nifQ
nifE
nifA
nifH
nifD
nifM
nifT
nifO
nifY
nifW
Phosphate solubilization
ACC deaminase activity
Urease activity
Siderophore (PGP)
Acetoin & butanediol synthesis
IAA production
Endotoxins
Capsules
Invasion
Siderophores (VF)
Adherence
Exotoxins
Ampicillin/Sulbactam
Trimethoprim/Sulfamethoxazole
Tetracycline
Piperacillin/Tazobactam
Gentamicin
Amikacin
Cefuroxime sodium
Ceftriaxone
Cefoxitin
Nitrofurantoin
Aztreonam
Cefazolin
Tobramycin
Cefepime
Ciprofloxacin
Ampicillin
Meropenem
Levofloxacin
Ceftazidime
Imipenem

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