Webinar Q&A

How genome-based classification can improve regulation of pathogens and beneficials

21 September 2021

Presenter: Boris Vinatzer, Professor, School of Plant and Environmental Sciences, Virginia Tech, USA

The webinar recording is available on the Phytobiomes Alliance YouTube channel at https://youtu.be/7kGsDA6fj2c

Q: If we classify too quickly taxonomically after a disease outbreak, wouldn't that cause issues as well with searches on e.g. early oubreak of COVD-19?

Timestamp: 43'40"

Q: For LIN (group or number) assignment, is there some bare minimum quality thresholds for input/query genome assembly?

Timestamp: 45'35"

Q: Do the LIN groups correspond to phylogenetic groups? What is the use/usefulness of phylogeny to the LIN base classification?

Timestamp: 47'00"

Q: For phenotyping, you would need to have accepted/validated techniques, right? I am thinking about phenotyping for pathogenicity or pathogenicity potential, for example.

Timestamp: 49'05"

Q: How do you define the first LIN number? To which specie is all the others compared?

Timestamp: 51'53"

Q: With respect to risk based on virulence genes, how does the LIN database take into consideration SNPs in key virulence genes that might cause shifts in phenotypes and subsequently affect regulatory decisions.

Timestamp: 53'33'

Q: How do you account for changes in host plant sensitivity - i.e via deployment of new cultivars - in risk assessment?

Timestamp: 55'10"

Q: Do you think LIN can replace scientific name in taxonomy?

Timestamp: 56'34"

Q: I can see this is a first step but I am uneasy about non-gene based classification. Small changes in genomes might well transform phenotype; horizontal transfer might transform a benign into a pathogenic strain, while a single SNP might transform a pathogen into a benign strain. Eukaryotic

microbes might well have divergent intergenic spaces while retaining the same pathogenic gene inventory.

Timestamp: 58'43"

Q: What happens in the case of convergent evolution, i.e. genetically distant strains and even species, that are pathogens to the same host?

Good question. Based on what I know, these pathogens may become more similar in regard to phenotype but not in regard to their genome sequences so they will be two separate groups that will be regulated separately.

Q: How to deal with phenotyping issue since a lot of proportion of new strains/variants are from developing countries where funding for this type of research is VERY limited

Yes, that is a huge challenge. I think we need much more international funding to investigate the phenotypic and genotypic diversity around the world.

Q: How to use Pseudomonas aerogenosa, a facultative pathogen as a biocontrol agent. What are the phenotyping methods to be adopted to rule out its human pathogenicity?

Good question. I think this article on the Burkholderiaceae presents a promising solution to this kind of question: https://www.mdpi.com/2073-4425/9/8/389

Q: how does the LINs system handle the presence of virulence genes that are not expressed?

Since the LIN system is exclusively based on genome sequence similarity, expression won't affect LIN assignment. Expression will affect phenotype of course but I am not aware of bacteria commonly maintaining virulence genes if they are not expressed.

Q: what about using virulence related primers to scale up the work with biological control burkholderia and bacillus

Yes, once specific virulence genes have been identified in a species or genus, their presence or absence based on PCR could speed up classification.

Q: Many pathogens have dozens of synonyms, and their current name may change again and again. How LinBase/genomeRiv adress such modifications? is it linked to database (like The catalogue of Life or Index fungorum) to name the groups/taxon accordingly?

Good question. For now, we have only applied LINs to bacteria but we are planning to include fungi in the future. Also, in our next version of LINbase, genomeRxiv, we will add both, the current NCBI taxIDs that are based on validly published names and the GTDB species clusters. Conceptually, it is easy to add any additional synonyms so users can see them all on the result pages. The question is that someone needs to do the work to enter the synonyms in the database. We will give subject experts access to genomeRxiv to do that.

Q: Could you elaborate on the LINgroup genome-independent primer design tool works? Will the PCR assays be specific to a LINgroup with minimal cross-amplification of other LINgroups?

That is the idea. My collaborator Leighton Pritchard has written a program that uses pyani results as input for the primer design. However, as is always the case with PCR primers, they need to be verified experimentally since it is impossible to predict how specific they are based on sequence comparisons alone.

Q: as a quick approach...usually genome sequence happened at advance stage of the projects

yes, that is still the case but I am convinced that genome sequencing will replace all other approaches in the future

Q: Are there specific genetic signatures in bacterial genomes that could help to predict potential bacterial lifestyle?

yes, good point. We are not there yet but the more we learn about more bacteria, the better we will become at predicting bacterial lifestyles based on their genome sequences. It will require a lot of phenotyping to gather there though.

Q: Which is the best omic tool to unravel the etiology of complex disease involving primary and secondary invaders?

That is clearly a challenge. I do not think there is a single tool that can effectively do that. Metagenomics can surely help reveal potential pathogens but interpreting the data in complex cases like you describe needs a lot expert knowledge. There is no simple solution.

Q: for the risk prediction, if applied to the crop field, we may need to integrate genomic, field disease phenotypes, microbiome, soil properties and environments to build machine leaning models or other models ? what your feeling of this application of disease prediction for farmers?

Yes, you are right. Risk depends on a lot of factors. The more we learn about all of them, the better we will become at predicting risk and then we can also leverage machine learning. But machine learning is only as good as the data. We need a lot of data to input to make machine learning become good at predicting risk based on so many different factors.

Q: how concerned are you about the effect of input order into the database, which may impact the classifications / LINgroups ?

Good question. We have not seen any significant effect on LINgroups based on input order. It may move the shared LIN positions one position to the left or right and that will affect the LINgroup circumscriptions slightly. However, we have not seen that causing a problem.

Q: Does the LIN approach have a protocol for dealing with changes in deposited sequences? For example, a better re-sequence, or the need to remove a mistaken genome?

Good question. When a genome of low quality is included in LINbase and a LINgroup that includes such a low quality genome is circumscribed, the breadth of the LINgroup may be affected and be too broad to allow precides identification. We had a case like this. We removed the low quality genome and then changed the LINgroup circumscription. We will try to avoid this as much as possible in the genomeRxiv by restricting the genomes we use to high quality genomes.

Q: phenotypic data may vary from lab to lab, as there might several factors, how LIN database will take care of errors in phenotype data, we can have quality parameters for genome data but can we have quality parameters for phenotypic data.

Excellent question. We need quality standards there as well. I forsee that most LINgroup circumscriptions will be based on phenotyping data described in peer-reviewed publications but if the peer-reviewed publication includes poor quality data then the LINgroup circumscription will be of poor quality.

Q: do you think that the pathovar concept is not appropriate to name the pathogens ?

It depends. Based on the research in my lab, there are some within-species phylogenetic groups that have specific host range based on field data and based on lab-based host range tests. In that case I think the phylogenetic group corresponds to a pathovar and such a pathovar could be describe as a LINgroup. It really depends on the quality of the host range data that are available.

Q: You said that LIN can reflect host range for pathogens that attack tomato or kiwi, for example. But has the host range of these pathogens been tested against other hosts even if they were not isolated from other hosts?

Good point. I feel most confident about host range if it is based on both on outbreaks that have occurred in the field and controlled host range tests. For example, the pathovar tomato lineage

that we called T1 in our publications has never been isolated from any outbreak other than bacterial speck of tomato outbreaks while the DC3000 lineage has been isolated from tomato and Brassicaceae disease outbreaks. In our controlled host range tests we obtained the same result. T1-lineage isolates did not cause disease on Brassicaceae but DC3000-lineage isolates did. In this case, I conclude that the two lineages do have a different host range from each other and I can circumscribe them separately in LINbase with their respective host ranges based on outbreaks in the field and controlled experiments. This may not always be the case though. There is no black and white.