Review manuscript# 122760v1

"Insights into antibiotic resistomes from metagenome-assembled genomes and gene catalogs of soil microbiota across environments" Han et al

This article presents data on over 1000 genomes assembled from 111 soil samples from multiple biomes from four different Chinese provinces, an impressive amount of work. It is a strength of this paper that there is a diversity of sampling in areas that are not well-represented in the databases. I think this is a useful scientific contribution. The methods and explanations of methods should be clarified in a number of places. Some discussion of possible implications would be welcome; the abstract promised insights but the paper reads more like a resource announcement than a study to inform the determinants of ARGs in various environments. Speculation about the correlations that they found would be fine, as long as they are identified as speculation.

21 Abstract "To investigate soil microorganisms in the areas where both humans and common domestic animals (such as pigs and
chickens) are present and active." Not a sentence.

The title refers to resistomes, which the authors don't mention much later. It is not well-defined and in one place seems to refer to four regional resistomes. At the end of the abstract, it seems to refer to a singular resistome. The introduction swings back and forth, too. And, the paper itself doesn't even use the term, as far as I could see.

48 soil is the origin of antibiotic resistance? Is that really established?

51 I don't think Chen et al argues that ARGs disrupt ecosystem function. Rather, there are some predictable patterns of pedogenesis, and soil age affects ARG abundance and diversity (generally in the positive direction).

52 awkward. Sentence implies exchange of resistance genes with humans

57 I wouldn't say that MAGs provide better resolution, but that they provide greater power.

59 Dove et al provides an example of the use of MAGs but it not a great choice to defend the statement being made. A better reference might be something like Anthony et al 2024 https://link.springer.com/article/10.1186/s40793-024-00599-w

71 This work might identify reservoirs but not vectors. Maybe "potential vectors"

78-79 There needs to be methods presented for sample storage, processing, next-gen sequencing, etc.

89 Sickle is really old. What was the reason for using it in addition to the other QC? Which data are the Sickle output?

96 also wondering about the use of the binning algorithms. MetaBAT2, MaxBin, and CONCOCT have a lot of overlap in binning functions. Are they each being tested ?

97 bins not binds? (Binning is also misspelled in Fig 1A)

100, 134 Has CheckM only been validated below 50% completeness? In Fig1B, there seem to be some genomes of 0.1mB, so unlikely to be complete. Lineage marker gene number? Source?

102 Is RefineM deprecated? https://github.com/donovan-h-parks/RefineM

113-how were these comparisons conducted? Manually? Prodigal doesn't call these databases, it just produces gff files of gene function.

114-123-Cite databases (e.g., CARD database https://card.mcmaster.ca/)

RESULTS

130 I'm not clear on the use of the multiple quality filtering algorithms. Were all three programs run sequentially on all reads? In Table S1, is the "clean reads" column following BWA host sequence removal and "optimized reads" following a different QC program? Which is which?

142-the phrase "heterogeneity in assembly fragmentation" is confusing to me. The heterogeneity doesn't likely arise from differences in DNA fragmentation in library prep, right? Are the differences due to variation in the success of the assembler software?

143-CDS prediction is done by just searching for ORFs?

145-wow, really pushing the upper limits on microbial GC content! Some large genomes here.

147-148 seems circular to me. The higher quality assemblies have a higher N50. What algorithm is this? How is the quality defined if not N50? Or ae you talking about quality as sequencing QC (PHRED scores)?

170- Figure 2B and 2C are very well-designed and attractive examples of data visualization! The legend for Fig 2A refers to only gray ribbons? What is the meaning of the color in the bands and ribbons?

193 what is the pipeline for this? I don't think it is explained in the Methods

226-"global overview maps"? What does that mean?

235- How is statistical significance being established here?

237-8-Is 4B a useful figure? There seems to be only minor differences by phylum.

249- Very minor nomenclature inconsistency: In the supplemental there is no 's' in the genome names. It makes more sense that way, I think, since the numbers refer to a single assembled genome. In the manuscript and the supplemental title, just list MAG817,MAG783, etc

249-Is there a cross-tabulation from sample to MAG? For example, for MAG817, can the reader find which sample that came from?

Fig S3- Can you readily remove the numbers from that heat map? What is the upper numerical boundary (what is the highest number of functional predictions?

265-66-This belongs in Discussion and could be expanded.

292-The correlation between MGEs and ARGs seems unremarkable. What are the points/lines along the axes? Total counts on that vertical or horizontal? I assume it is not zero values for one variable or the other?

What is the x-axis for Figure 5C? Just the number of all protein coding sequences??

295-The correlation between transposase genes and ARGs is expected, so it is good to see here. Any idea about the negative correlation between the ATPase and ARGs? TraG is a protein in secretion, right? Some speculation about this correlation could be made in Discussion.

DISCUSSION

313-I'd like more literature here comparing the results to other areas. Are there differences in representation compared to other global sampling? Are archaea overrepresented in this sample? Provide some context rather than just repeating Results.

318-state more specifically what the environmental conditions are like in the sampling site in Yunnan province. FIgS5 can be discussed in terms of which environmental conditions are correlated with increased ARG representation. For readers unfamiliar with the natural environment at the sampling sites, can some information about environment (from table S1) be summarized and compared to abundance of ARGs. [Given the prominence of ARGs in the Title and Abstract there is a good argument for including FigS5 and Fig S6 in the main body of the paper. Fig 1B and 1C, especially 1B seem more tangential and could perhaps go into Supplemental.]

332- In line 214, the authors mentioned that two classes of carbohydrate enzymes were notable. What are the implications of that enrichment? What is the context for that finding? Likewise, is it expected that those genera mentioned in 227 would have high proportions of carbohydrate encoding genes?

333-Having a genome primarily concerned with metabolism seems obvious to me. Is there anything about the functional annotation that was interesting or unexpected? If not, the discussion could still serve to frame the functional ontology in the context of expected findings/previous studies.

346-This is the first mention that these are pristine environments for ARGs, I think. Have the investigators established this fact? In particular, do we know whether or not the farms are using antibiotics? I think this fact could change the direction of the paper markedly.

353-There is no mention of the archaea that contain ARGS; Strikingly, *Methanobacteriota* has representatives from almost every class of ARGs! What is your interpretation of this? Are these under pressure or did they evolve these mechanisms just under natural conditions?

359-There was also a correlation between transposases and ARGs, which strikes me as quite intuitive that transposase activity would encourage horizontal transfer of these genes. It is intuitive but worth commenting on since it is a result that might be expected, thus helping to validate the method.

CONCLUSIONS

381- this just repeats a point that was in the discussion and is too vague to be useful.

394-I'd be more interested in environmental determinants of the ARG prevalence more than just increasing sample size.

Authors- no one is listed for sample collection and processing?

Data availability – it doesn't appear that the authors are required by funding for making these data publicly available, so they are to be commended for doing so. In addition to the BIoProject number, they could include the SRA numbers: SRX22037149- SRX22037259

Table S10 Is MajorBIoCloud a mirror of the Korean EzBioCloud?

Editor request: "Field study-Have you checked the authors field study permits? Are the field study permits appropriate?"

I am not well-equipped to assess the appropriateness of Chinese field study permits.