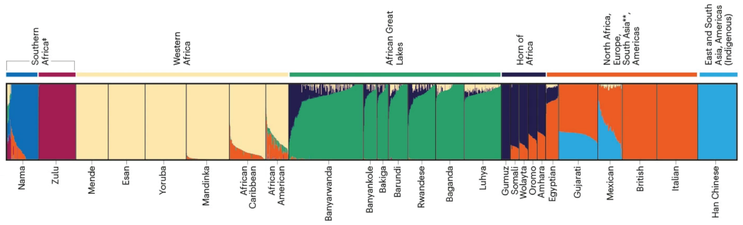
Structural variant discovery and characterization from *de novo* assembly of Khoe-Sān genomes

Zoeb N. Jamal[1*], Daniela C. Soto[1*], Kristin Hardy[1*], Mohamed Abuelanin[1], William Palmer[1], Javier Prado-Martinez[2], Paul Norman[3], Marlo Moller[4], Brenna M. Henn[1], Megan Y. Dennis[1]

[1]Genome Center, University of California, Davis, CA. [2]Institute of Evolutionary Biology, PRBB, Barcelona, Spain.
[3]University of Colorado, Biomedical Informatics, Aurora, CO. [4]Stellenbosch University, Molecular Biology and Human Genetics, Stellenbosch, South Africa.

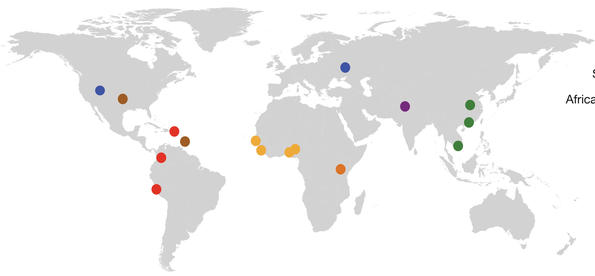
Background

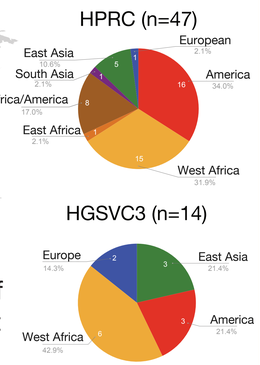
- Long read sequencing (LRS) of large cohorts is enhancing our understanding of the genomic variation landscape
- Genetic diversity within African populations can exceed that between continental groups, yet many remain underrepresented



Carlson et al. 2022 PMID: 36261568

- This gap suggests ample genetic diversity is missing in large datasets

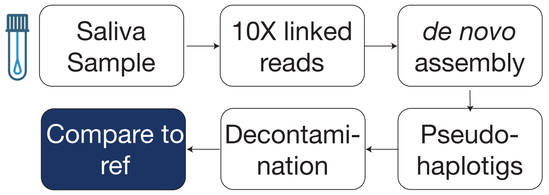


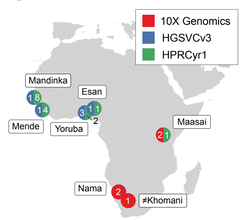


- In particular, the Khoe-Sān indigenous peoples of Southern Africa carry some of the most divergent haplotypes in extant human lineages

Methods

- We performed 10X genomics long range sequencing and de novo assembly for 3 Khoe-Sān individuals (2 Nama, 1 ≠Khomani San)

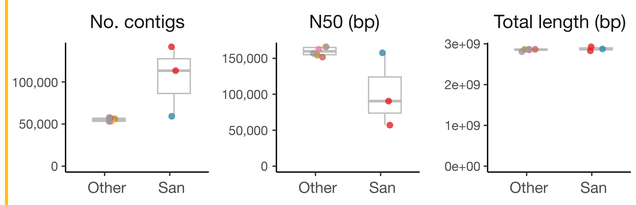




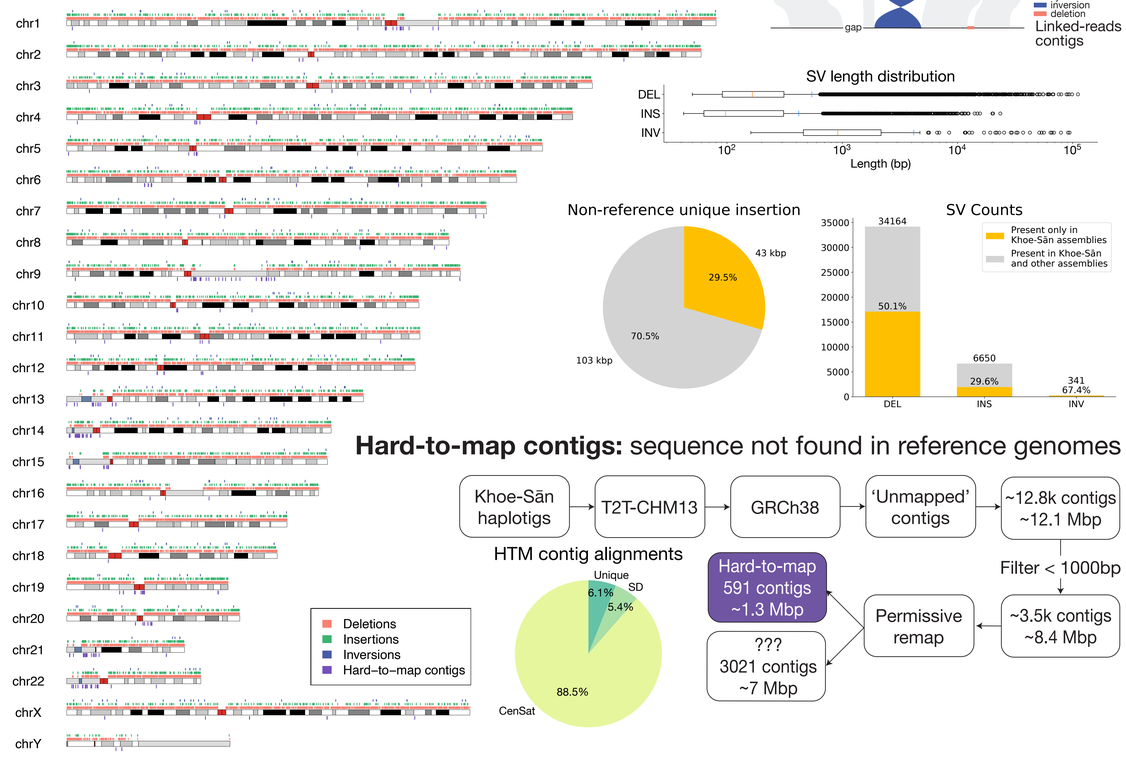
- We called structural variants (SVs) against T2T-CHM13 for these Khoe-Sān genomes and other 10X linked-read assemblies
- Used whole genome short read sequencing (SRS) data from a cohort of 93 Khoe-Sān individuals to genotype discovered SVs and investigate copy number variability
- Used VEP to query the impact of genotyped SVs

Results

Assembly statistics: *de novo* assembly compared to other 10X linked read assemblies

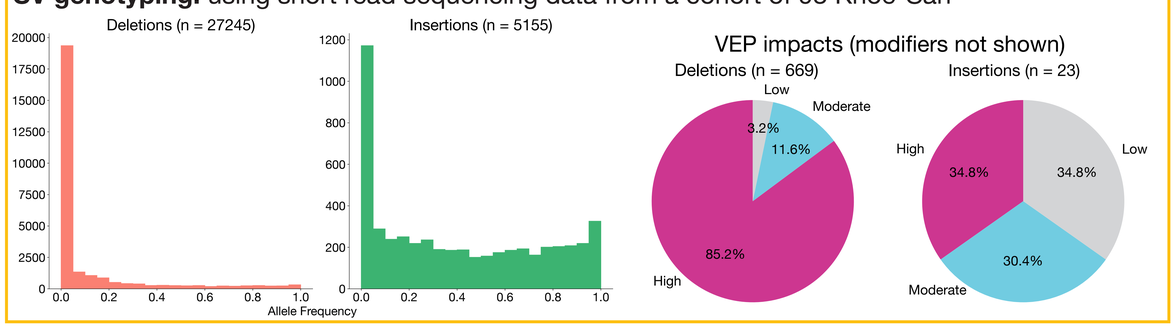


Structural variation: assembly to reference comparison

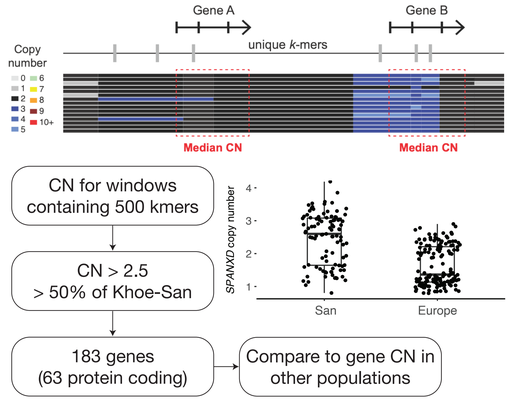


Hard-to-map contigs: sequence not found in reference genomes

SV genotyping: using short read sequencing data from a cohort of 93 Khoe-Sān



Copy number variation: using short reads from 93 Khoe-Sān individuals



Conclusions

- A lack of diversity in African genomes in LRS projects limits us from exploring human genetic variation
- We generated 10X linked-read and de novo assembly for 3 Khoe-Sān individuals
- We found ~7Mbp not mapping to human reference genomes, even after relaxing parameters
- We ascertained allele frequencies of many novel structural variants through genotyping in SRS Khoe-Sān samples
- We detected copy number variations that may differ from other human populations

Future directions

- Robust investigation of non-mapping contigs, namely aligning short read Khoe-Sān datasets to them
- Genotype discovered SVs in other populations from the HGDP, and identify Khoe-Sān specific variants
- Calculate F_{ST} and V_{ST} between Khoe-Sān and HGDP to study stratification
- Further explore high impact variants highlighted by VEP
- Leverage long haplotypes to investigate structurally complex loci, including HLA
- Assess improved mappability for Khoe-Sān individuals using new assemblies

References

Carlson et al. 2022 PMID: 36261568, Chen et al. 2019 PMID: 31856913
Shen and Kidd. 2020 PMID: 32013076

Acknowledgements

Thank you to funding sources (NIMH: R01MH132818, NIGMS: R35GM133531) and collaborators for their help facilitating this project, as well as the HGDP and HPRC for access to their data.

UC DAVIS SCHOOL OF MEDICINE
UC DAVIS GENOME CENTER
NIH National Institutes of Health