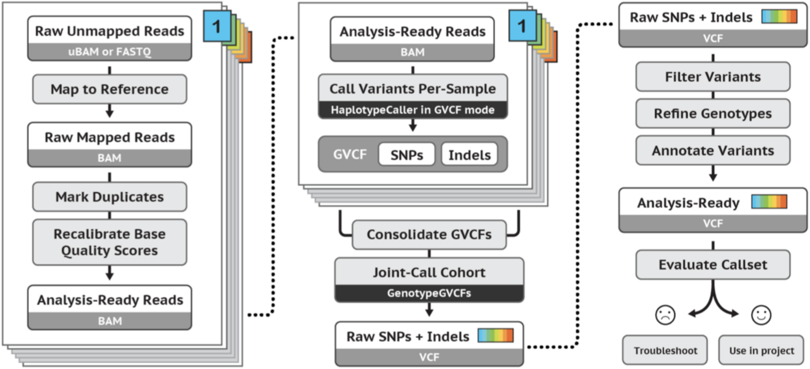
Draft of a Saudi pangenome graph

Robert Hoehndorf

KAUST Center of Excellence for Smart Health (KCSH)
King Abdullah University of Science and Technology

PacBio Day @ KAUST (9 October 2024)

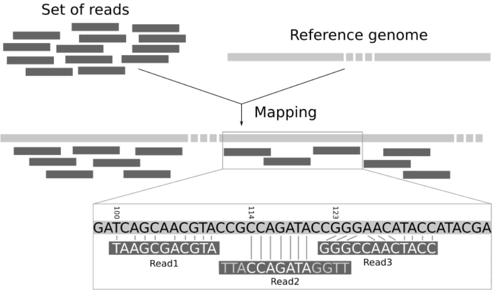
Variant calling



Variant calling is (still) challenging

- compute
  - around 30 USD worth compute time on cloud per genome
  - 12-24 hours compute
- storage
  - up to 1 TB for a trio
- resources: prior knowledge on known variants (population-specific), reference genome
- structural variants, chromosomal abnormalities
- reproducibility and provenance
  - many possible variations of a workflow
  - differences in program/data version

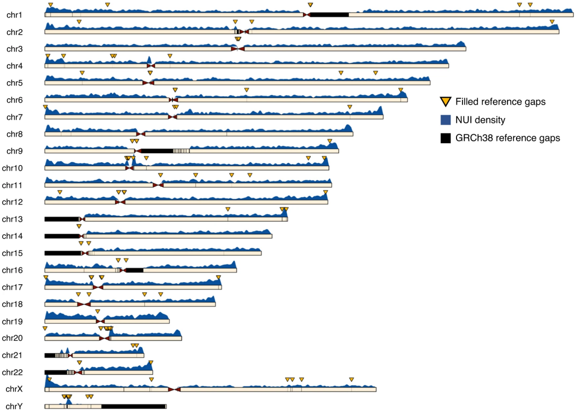
Reference genome



- linear, haploid
- single individual or small group of individuals
- major allele references

The use of single, haploid, linear reference genomes leads to reference bias

Reference bias: gaps and missing structural variation



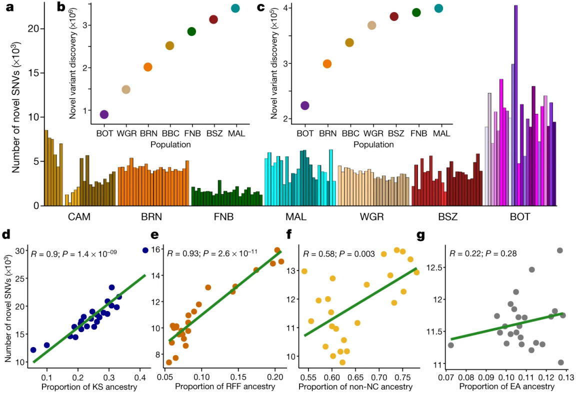
▽ Filled reference gaps

🟦 NUI density

■ GRCh38 reference gaps

Wong et al., Nat Comm 2020

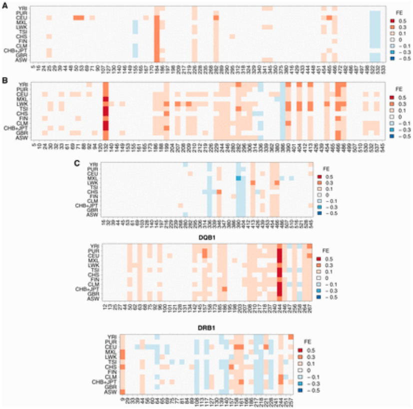
Reference bias: lack of diversity



Choudhury et al., Nature 2020

- lack of diversity
  - greater problem as distance to reference increases
  - structural variation
  - reads won’t map

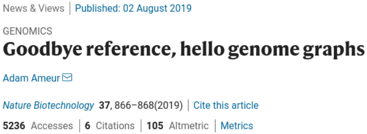
Reference bias: effect on genomic medicine



- Brandt et al. (G3 2015) compared 1000 Genomes Phase 1 with Sanger sequencing of HLA-A, -B, -C, -DRB1, and -DQB1
- 18.6% of HLA genotypes were mismatched
- usually: overestimation of reference allele (mapping bias)
- affecting any polymorphic region

Improving variant calling

- does variant calling need a reference genome?



News & Views | Published: 02 August 2019

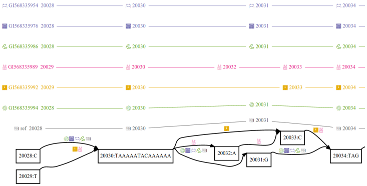
GENOMICS

Goodbye reference, hello genome graphs

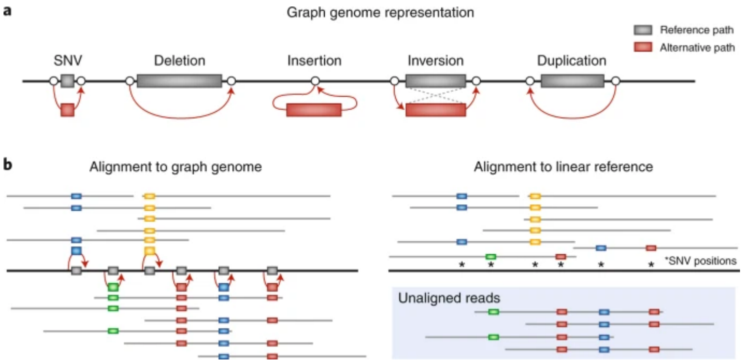
Adam Ameur

Nature Biotechnology **37**, 866–868(2019) | Cite this article

5236 Accesses | **6** Citations | **105** Altmetric | Metrics

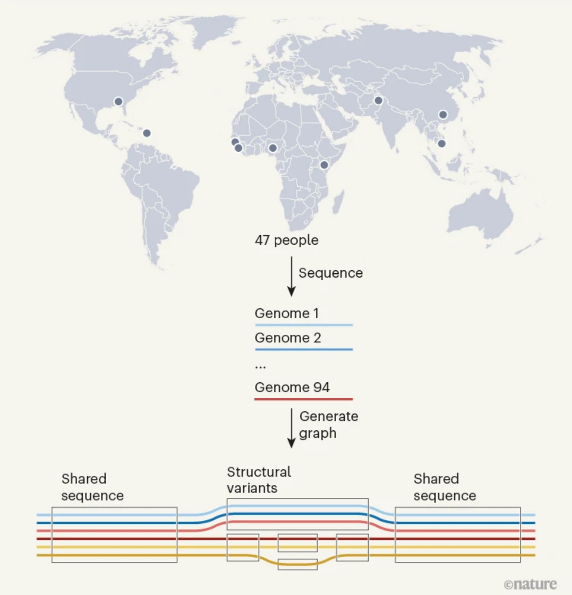


Improving variant calling: Pangenome graphs

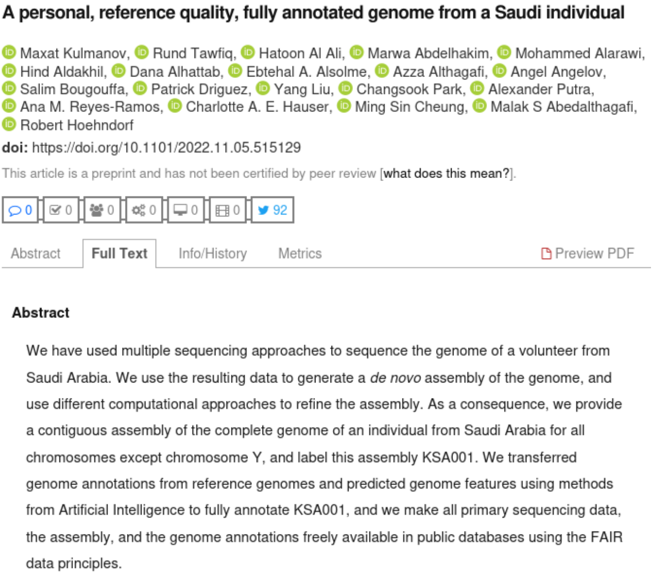


From Nat Biotechnol 37, 866–868 (2019)

HPRC graph lacks Arab population



KSA001: T2T Saudi genome



A personal, reference quality, fully annotated genome from a Saudi individual

Maxat Kulmanov, Rund Tawfiq, Hatoon Al Ali, Marwa Abdelhakim, Mohammed Alarawi, Hind Aldakhil, Dana Alhattab, Ebtehal A. Alsolme, Azza Althagafi, Angel Angelov, Salim Bougouffa, Patrick Driguez, Yang Liu, Changsook Park, Alexander Putra, Ana M. Reyes-Ramos, Charlotte A. E. Hauser, Ming Sin Cheung, Malak S Abedalthagafi, Robert Hoehndorf

doi: https://doi.org/10.1101/2022.11.05.515129

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

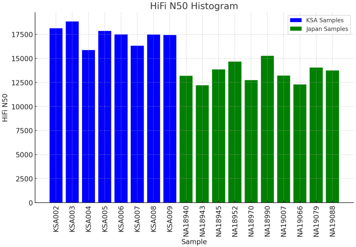
We have used multiple sequencing approaches to sequence the genome of a volunteer from Saudi Arabia. We use the resulting data to generate a *de novo* assembly of the genome, and use different computational approaches to refine the assembly. As a consequence, we provide a contiguous assembly of the complete genome of an individual from Saudi Arabia for all chromosomes except chromosome Y, and label this assembly KSA001. We transferred genome annotations from reference genomes and predicted genome features using methods from Artificial Intelligence to fully annotate KSA001, and we make all primary sequencing data, the assembly, and the genome annotations freely available in public databases using the FAIR data principles.

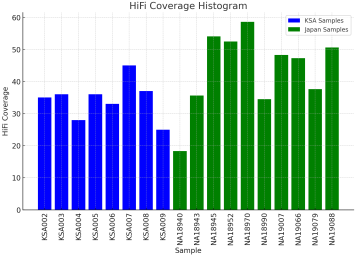
Japanese–Saudi PanGenome (JaSaPaGe)

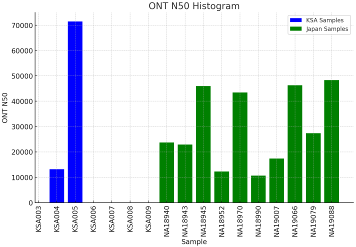
- aim: population-specific pangenome
- 9 Saudi volunteers (7 male, 2 female)
- 10 Japanese samples (Coriell, all male)
- sequencing:
  - Pacbio HiFi: 30-50x coverage
  - Hi-C: 50x coverage
  - ONT (UL): 7-100x

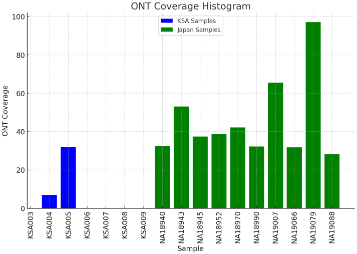
Unpublished results

Sequencing statistics



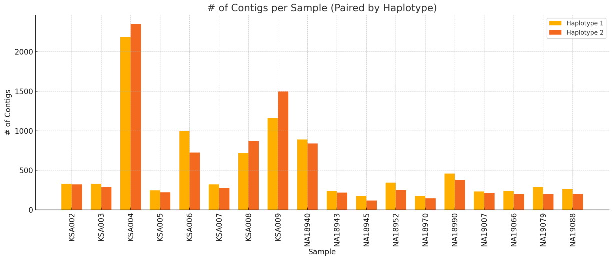




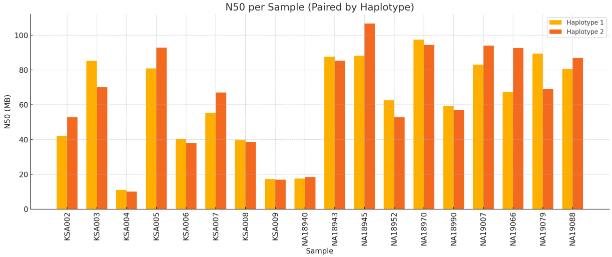


Unpublished results

Genome assembly



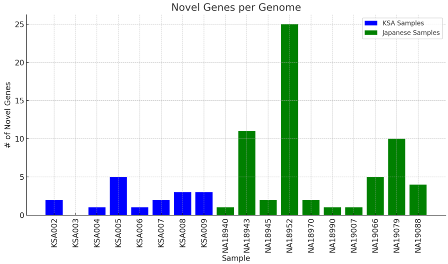
# of Contigs per Sample (Paired by Haplotype)

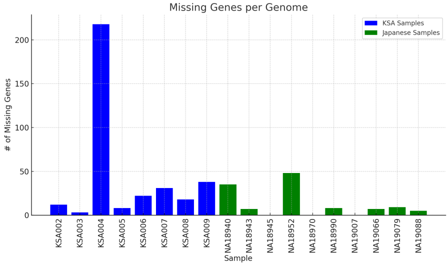


N50 per Sample (Paired by Haplotype)

Unpublished results

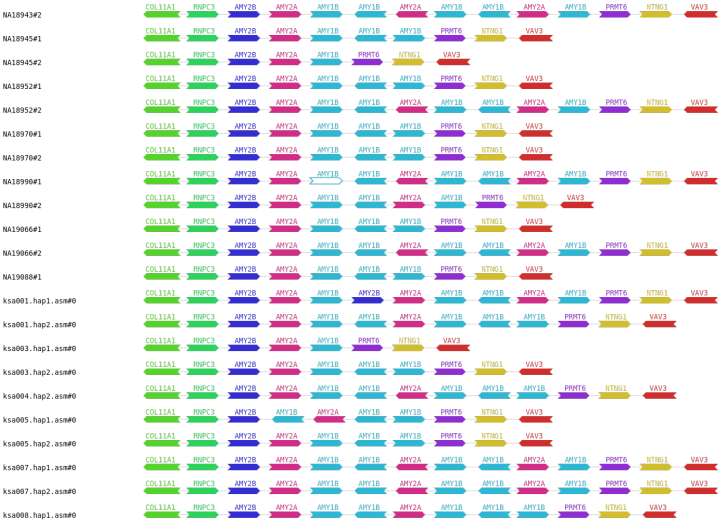
Genome assembly





Unpublished results

Copy number variants, Amylase



NA18943#2 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY2A AMY1B PRMT6 NTNG1 VAV3

NA18945#1 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

NA18945#2 COL11A1 RNPC3 AMY2B AMY2A AMY1B PRMT6 NTNG1 VAV3

NA18952#1 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

NA18952#2 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY2A AMY1B PRMT6 NTNG1 VAV3

NA18970#1 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

NA18970#2 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

NA18990#1 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY2A AMY1B PRMT6 NTNG1 VAV3

NA18990#2 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B PRMT6 NTNG1 VAV3

NA19066#1 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

NA19066#2 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY2A AMY1B PRMT6 NTNG1 VAV3

NA19088#1 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

ksa001.hap1.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B PRMT6 NTNG1 VAV3

ksa001.hap2.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

ksa003.hap1.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B PRMT6 NTNG1 VAV3

ksa003.hap2.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

ksa004.hap2.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

ksa005.hap1.asm#0 COL11A1 RNPC3 AMY2B AMY1B AMY2A AMY1B AMY1B PRMT6 NTNG1 VAV3

ksa005.hap2.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

ksa007.hap1.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY2A AMY1B PRMT6 NTNG1 VAV3

ksa007.hap2.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY2A AMY1B PRMT6 NTNG1 VAV3

ksa008.hap1.asm#0 COL11A1 RNPC3 AMY2B AMY2A AMY1B AMY1B AMY2A AMY1B AMY1B AMY1B PRMT6 NTNG1 VAV3

Unpublished results

Availability

- Creative Commons Zero (Public Domain)
  - use for any purpose
- `https://github.com/bio-ontology-research-group/ksa001`
- NCBI Datasets
- Sequence Read Archive
- all data available by November
  - missing: pangenome graphs, assemblies

Unpublished results

Acknowledgements

- Maxat Kulmanov (KAUST)
- Malak Althagafi (Emory / KFMC)
- Yosuke Kawai (NCGM)
- Toshiaki Katayama (DBCLS / DDBJ)
- Marwa Abdelhakim (KAUST)
- Saeideh Ashouri (NCGM)
- Yang Liu (KAUST)
- Wei Ran (NCGM)
- Rund Tawfiq (KAUST)
- Bioscience Corelab (KAUST)

Funding: Smart Health Initiative, KAUST Center of Excellence for Smart Health

Thank you