Genotyping structural variation

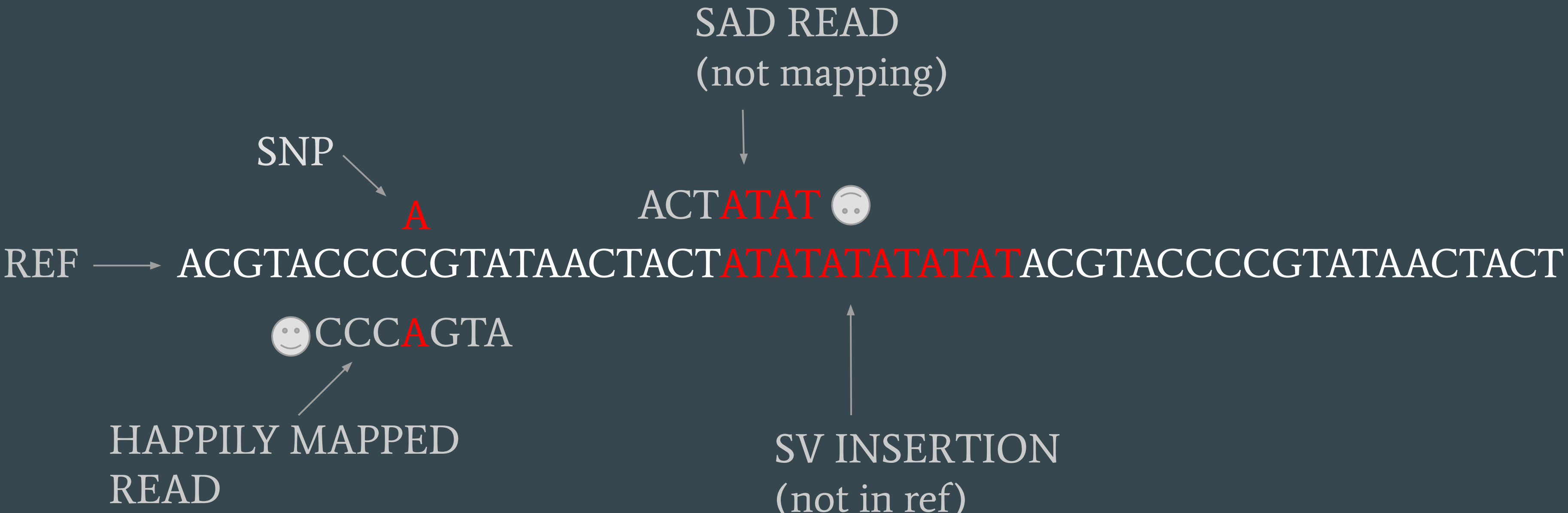
In the era of pangenomes

Ivar Grytten - Norwegian Biodiversity and Genomics Conference 2024

Genotyping structural variation

1. The challenge of genotyping structural variation
2. The role of pangenomes
3. KAGE

Genotyping structural variation is trick~~y~~ier than SNPs/indels



SAD READ
(not mapping)

SNP

A

ACTATAT

REF ACGTACCCCGTATAACTACTATATATATATATATACGTACCCCGTATAACTACT

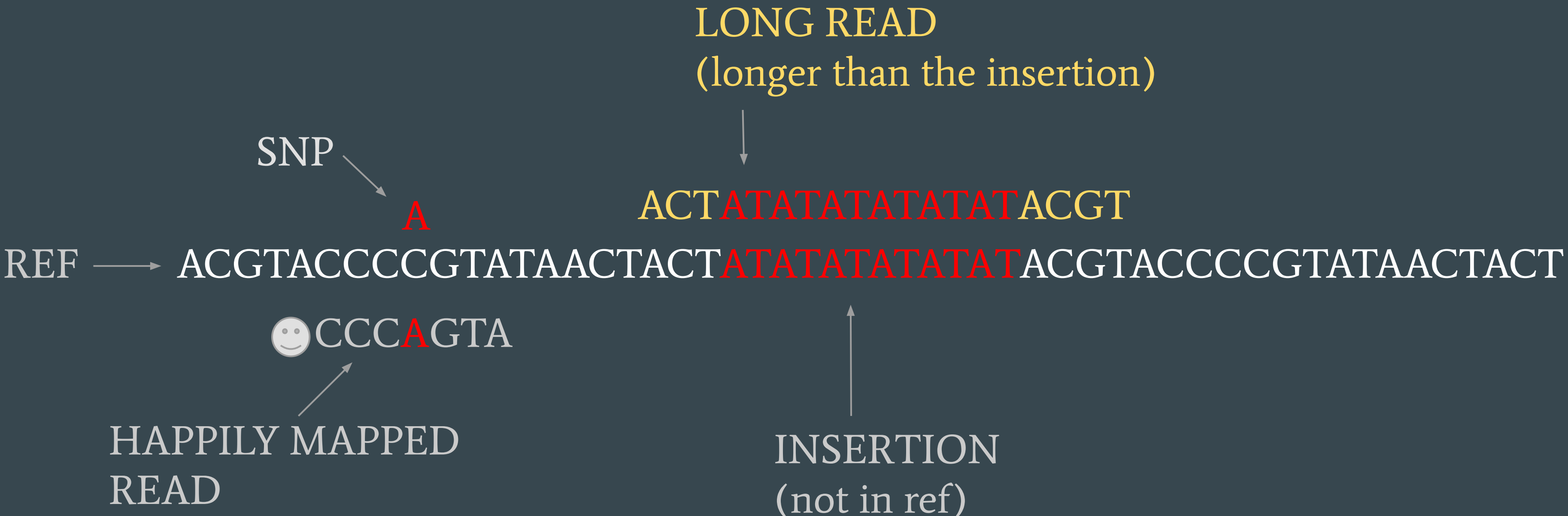
CCCAGTA

HAPPILY MAPPED
READ

SV INSERTION
(not in ref)

... mapping short reads to a reference genomes is bad for detecting SVs

Long reads can solve this



LONG READ
(longer than the insertion)

SNP

ACTATATATATATATACGT

A

REF ACGTACCCCGTATAACTACTATATATATATATACGTACCCCGTATAACTACT

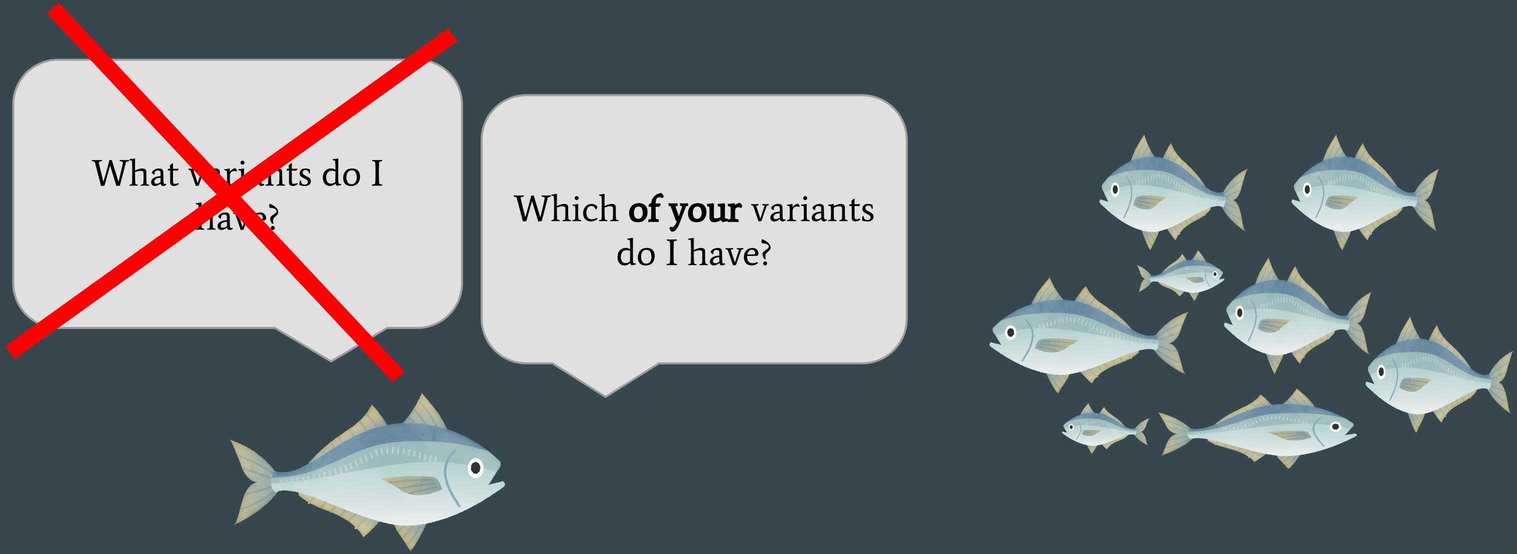
☺CCCAGTA

HAPPILY MAPPED
READ

INSERTION
(not in ref)

.. but long reads are expensive

Pangenomes are changing how we “call variants”



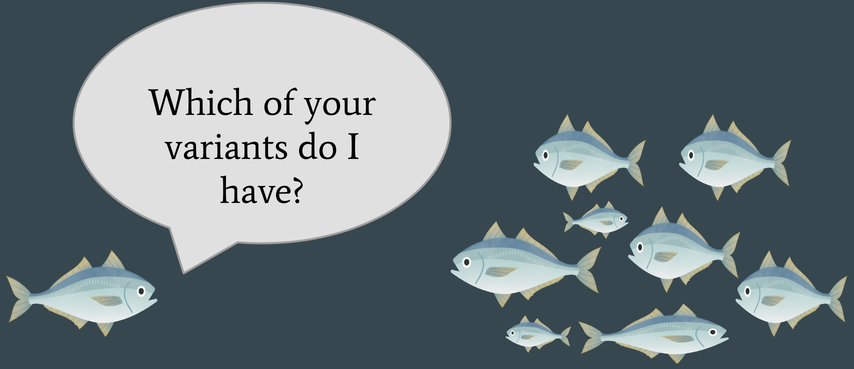
~~What variants do I have?~~

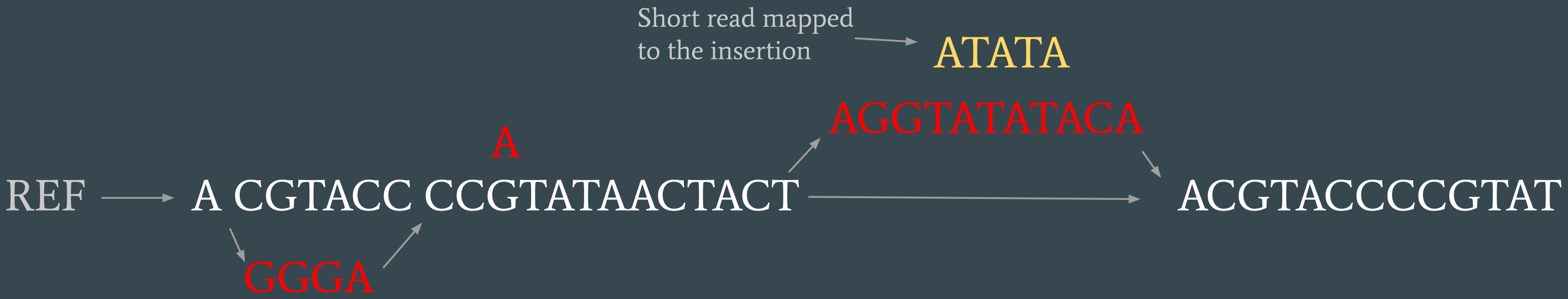
Which **of your** variants do I have?

Pangenomes are changing how we “variant call” / genotype

If we know the variation present in a population:

call sample by *genotyping known variants*





Short read mapped to the insertion

ATATA

AGGTATATACA

A

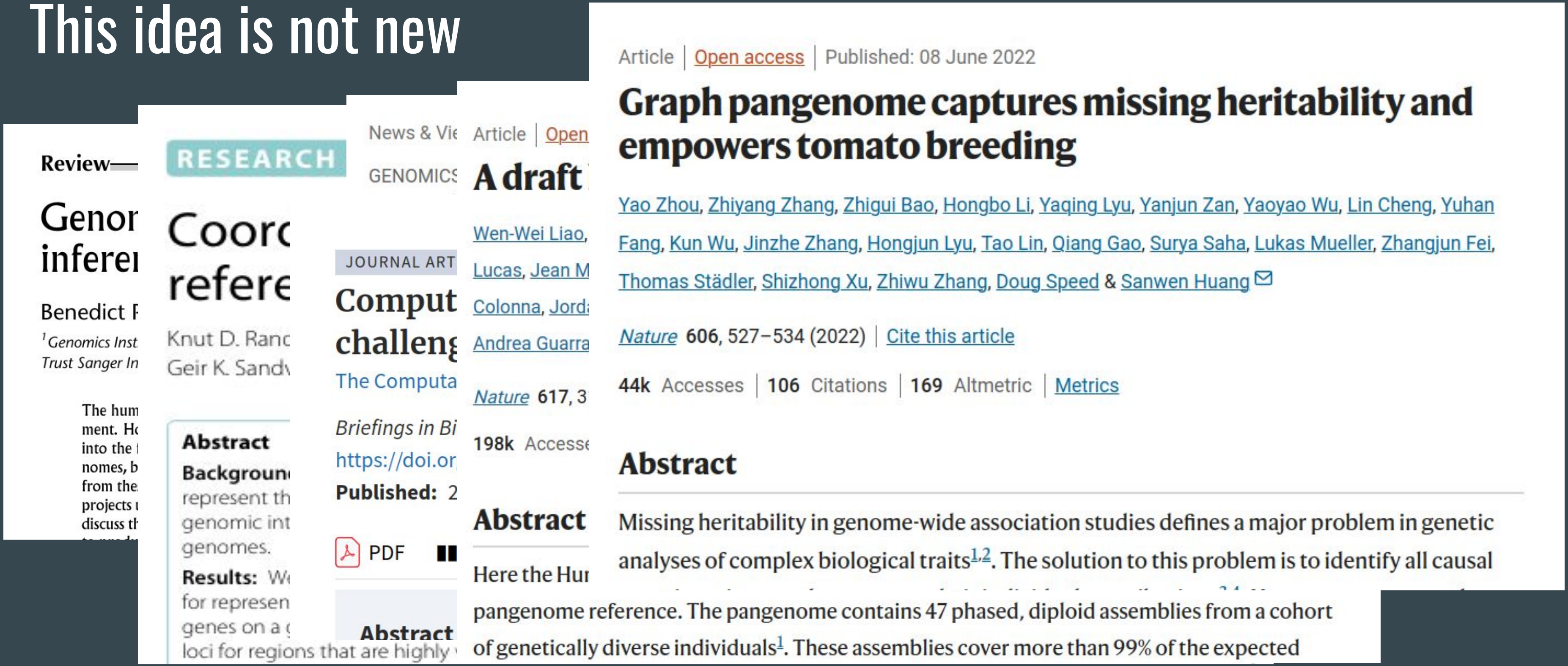
REF

A CGTACC CCGTATAACTACT

ACGTACCCCGTAT

GGGA

This idea is not new



Review—

Genor
inferei

Benedict F

[1] Genomics Inst
Trust Sanger In

The hum
ment. Ho
into the
nomes, b
from the
projects
discuss th

RESEARCH

Coord
refere

Knut D. Ranc
Geir K. Sandv

Abstract

Background: represent th
genomic int
genomes.

Results: We
for represen
genes on a
loci for regions that are highly

News & Vie

GENOMICS

JOURNAL ART

Comput
challeng

The Computa

Briefings in Bi
https://doi.or
Published: 2

PDF

Abstract

Article | Open

A draft

Wen-Wei Liao,
Lucas, Jean M
Colonna, Jord
Andrea Guarra

Nature 617, 3

198k Accesse

Abstract

Here the Hur
pangenome reference. The pangenome contains 47 phased, diploid assemblies from a cohort of genetically diverse individuals[1]. These assemblies cover more than 99% of the expected

Article | Open access | Published: 08 June 2022

Graph pangenome captures missing heritability and empowers tomato breeding

Yao Zhou, Zhiyang Zhang, Zhigui Bao, Hongbo Li, Yaqing Lyu, Yanjun Zan, Yaoyao Wu, Lin Cheng, Yuhan Fang, Kun Wu, Jinzhe Zhang, Hongjun Lyu, Tao Lin, Qiang Gao, Surya Saha, Lukas Mueller, Zhangjun Fei, Thomas Städler, Shizhong Xu, Zhiwu Zhang, Doug Speed & Sanwen Huang ✉

Nature 606, 527–534 (2022) | Cite this article

44k Accesses | 106 Citations | 169 Altmetric | Metrics

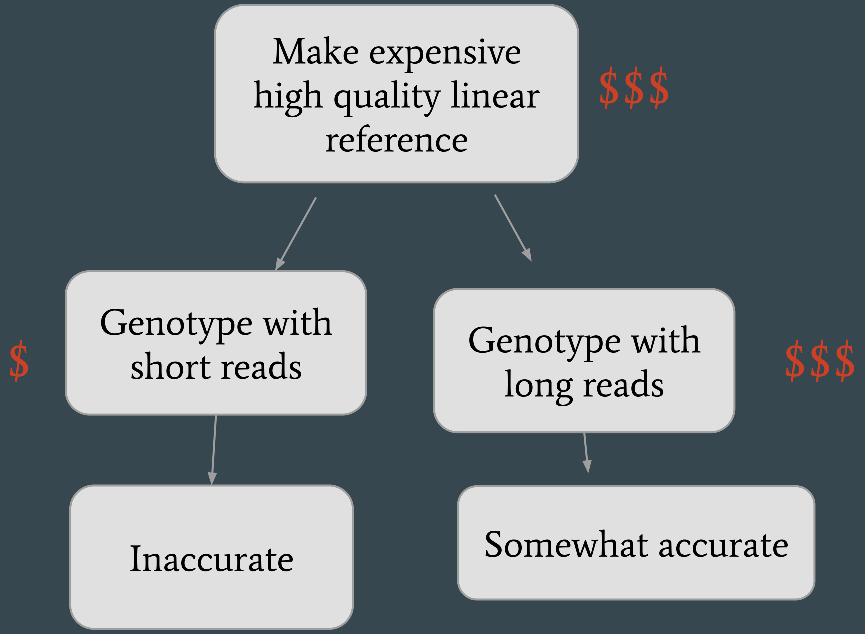
Abstract

Missing heritability in genome-wide association studies defines a major problem in genetic analyses of complex biological traits[1,2]. The solution to this problem is to identify all causal

.. but good assemblies are making it relevant now

The pangenomic approach to genotyping

Traditional approach



Make expensive high quality linear reference $$$

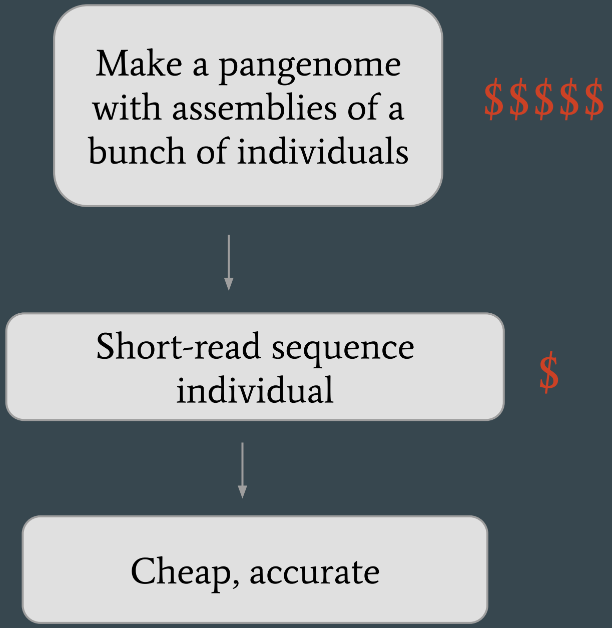
$ Genotype with short reads

Genotype with long reads $$$

Inaccurate

Somewhat accurate

Pangenome approach



Make a pangenome with assemblies of a bunch of individuals $$$$$

Short-read sequence individual $

Cheap, accurate

.. we have the assemblies, but what about the tools?

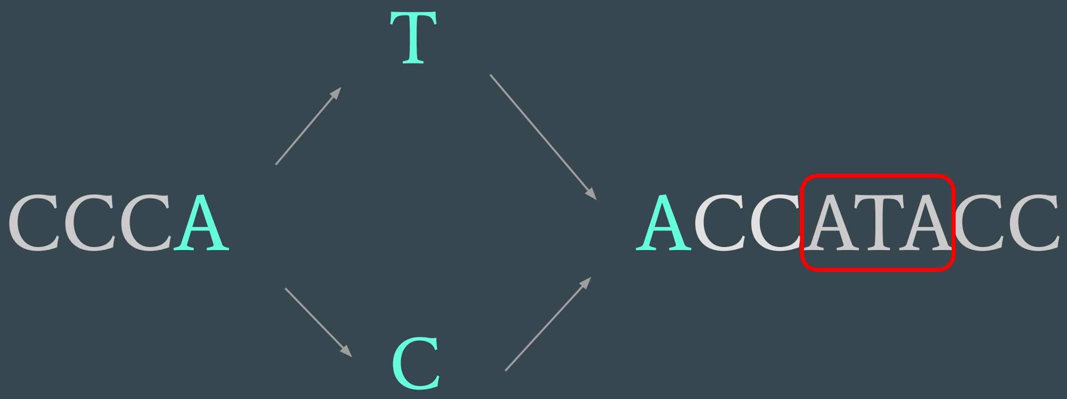
KAGE enables fast and accurate genotyping using pangenomes

- KAGE uses a graph-representation of known variants in a population
- Alignment-free, only looks at kmers (fast)
- **Two** key novel ideas makes KAGE pretty good

A puzzle: Which genotype?

We sequence an individual and get these 4 reads:

ATA
ATA
ACA
ACA



T

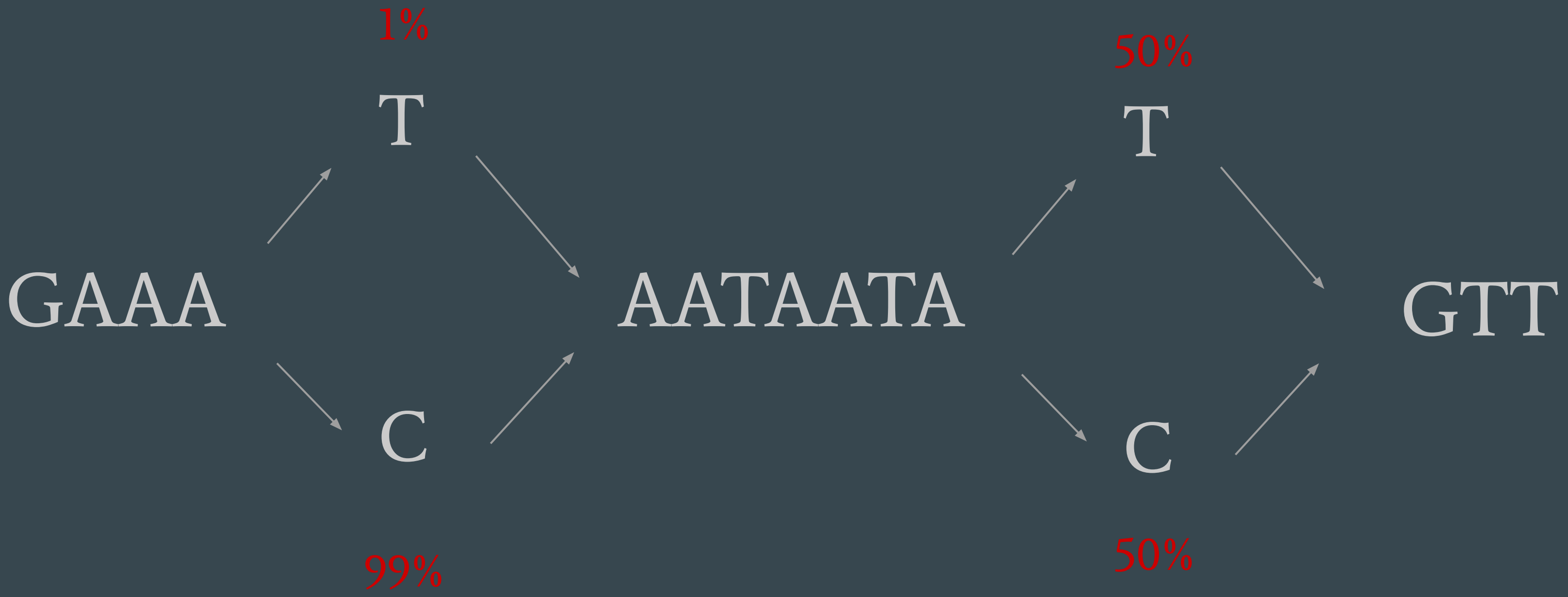
CCCA

ACCATACC

C

The ATA supports the variant, but we expect higher ATA-count due to the repeat.

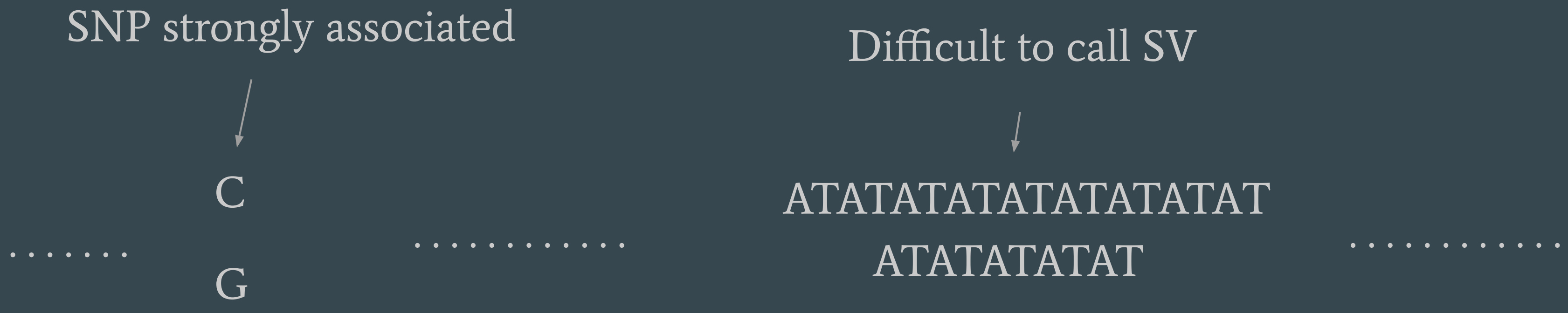
It helps to look at multiple variants together



P(T | C at other variant) = 0.9999

Non-unique kmers and repetitive sequences are common for SVs

- Since KAGE models these, we can genotype variants that are otherwise tricky



SNP strongly associated

C

G

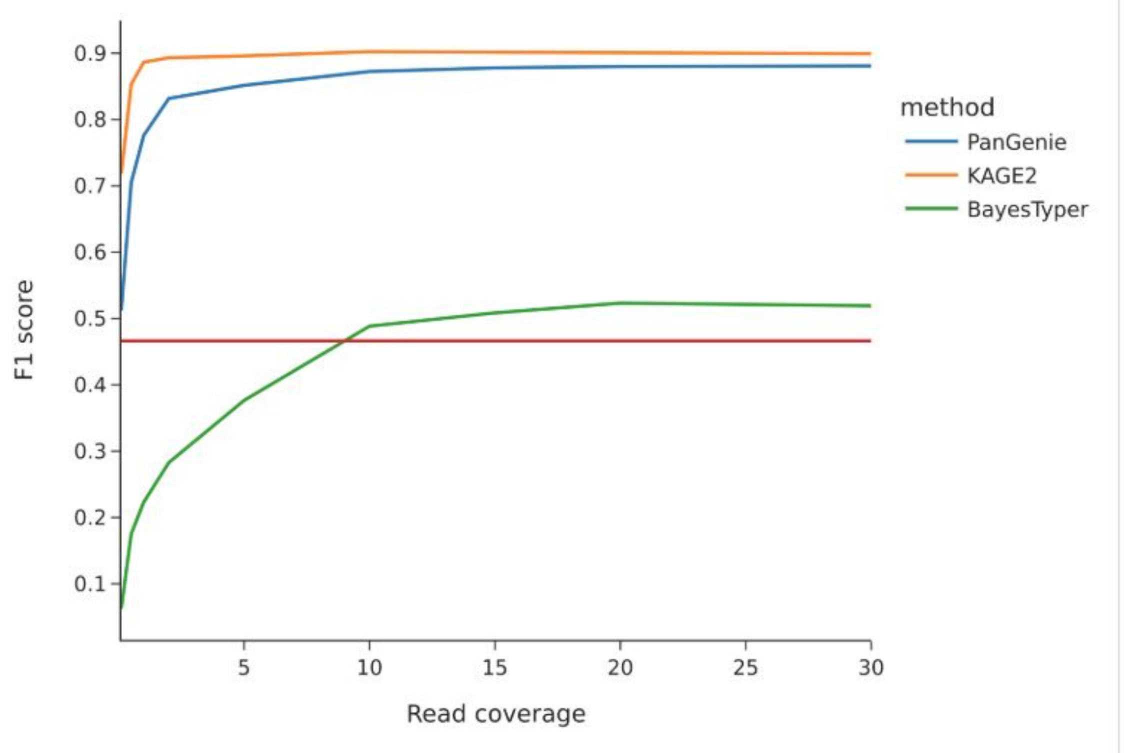
Difficult to call SV

ATATATATATATATATAT

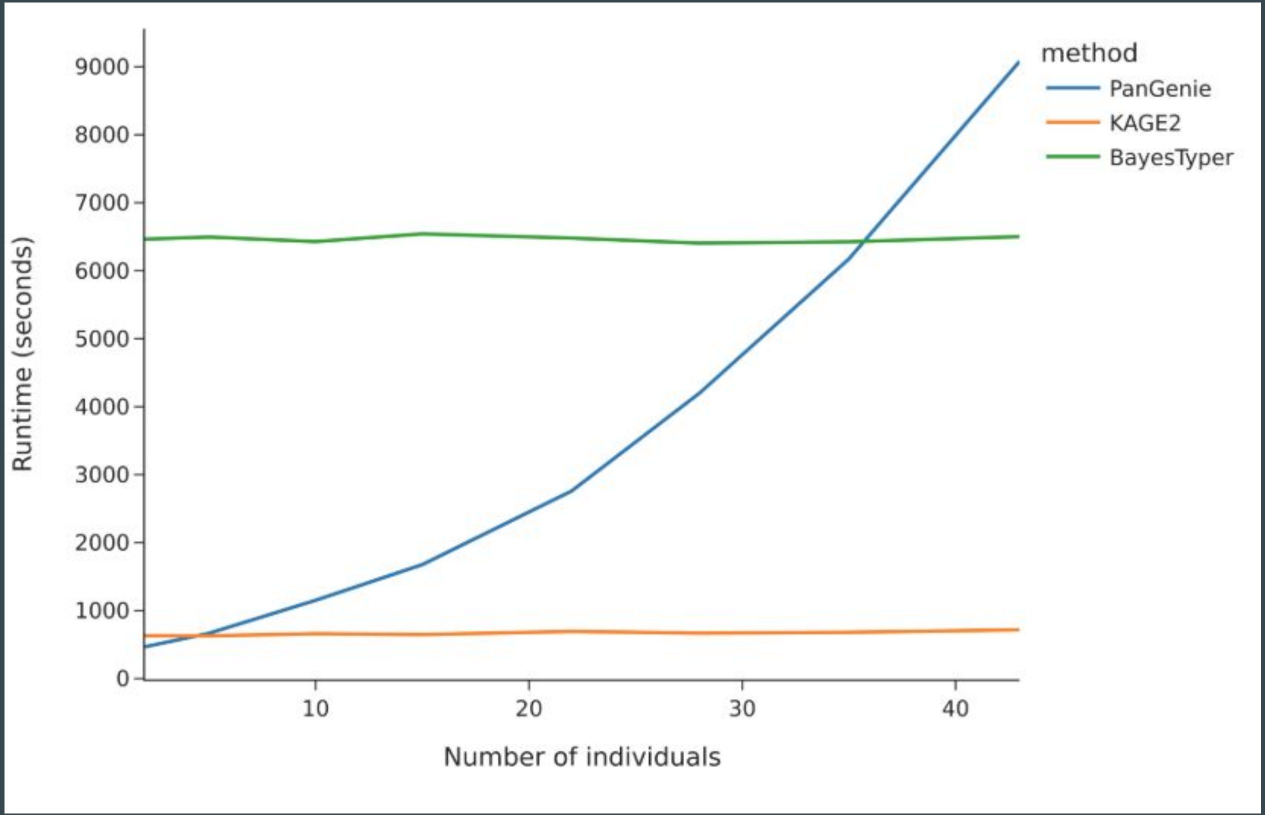
ATATATATAT

Easy-to-genotype SNPs and indels guide SV-genotyping

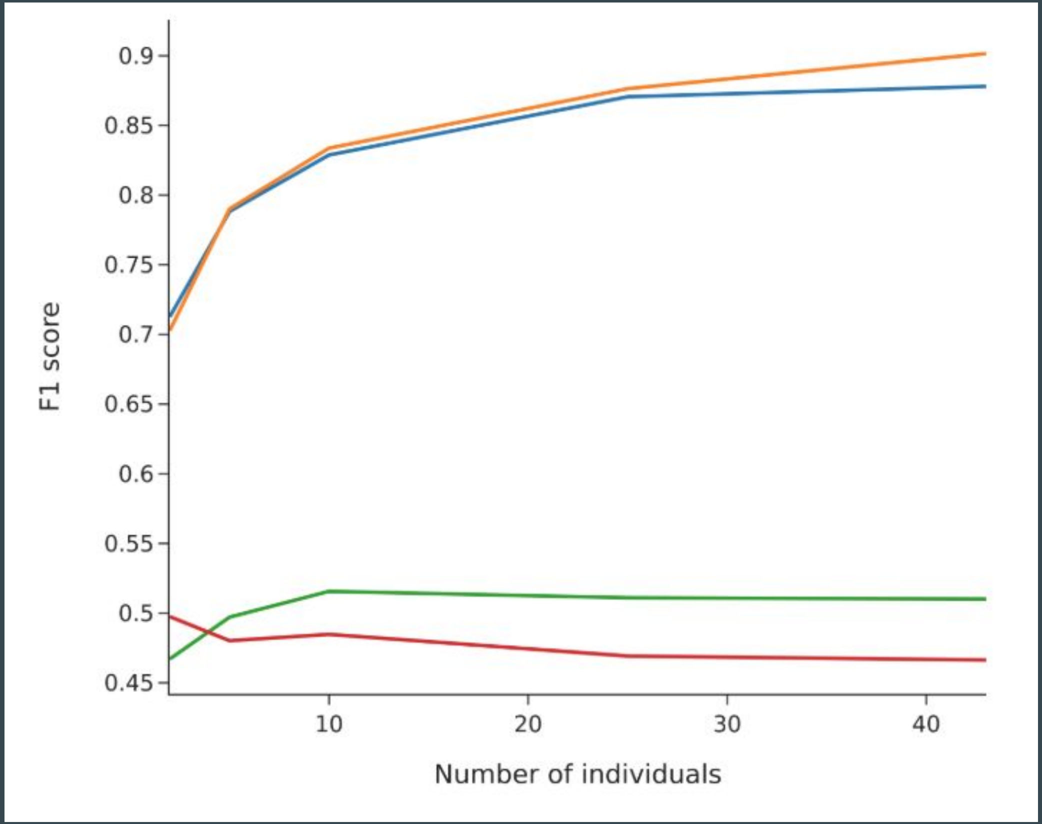
Good accuracy even when low read-coverage



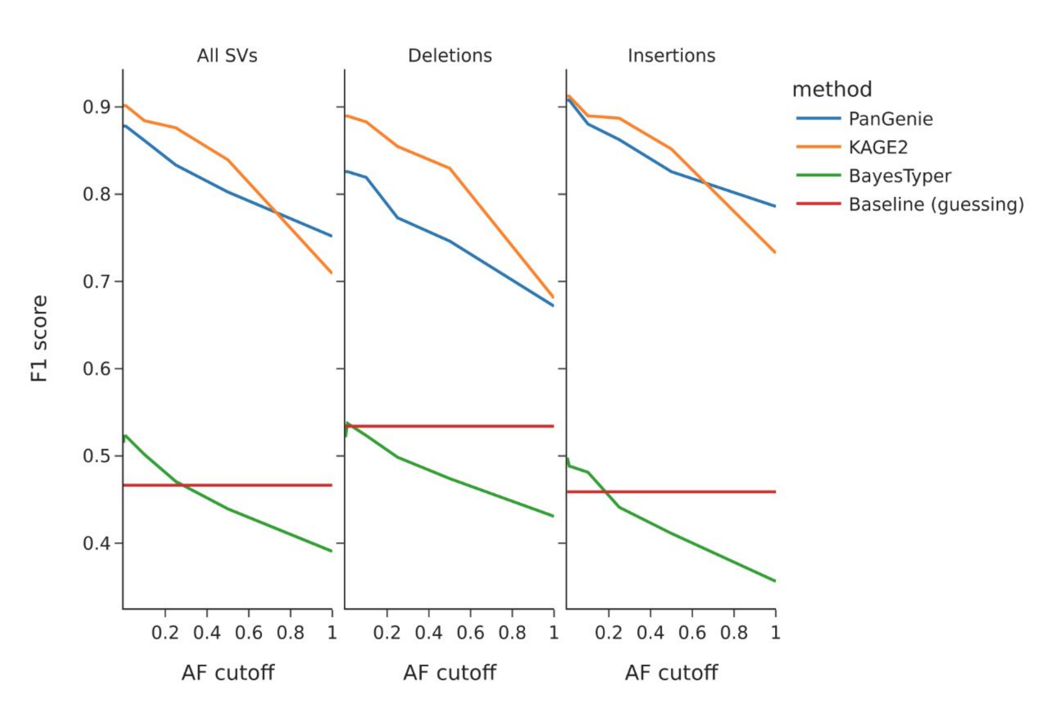
KAGE scales well to LARGE pangenomes



Larger pangenomes: Higher accuracy



SNPs and Indels help SV-genotyping

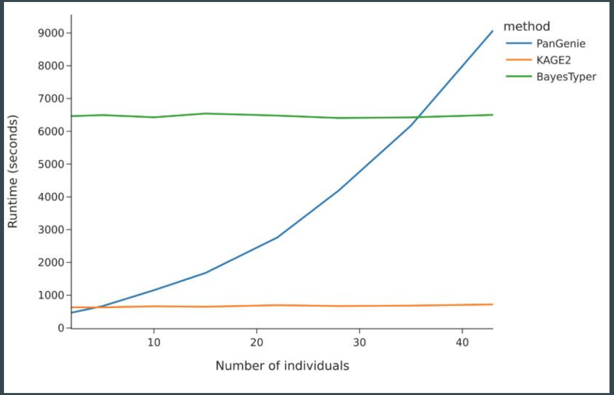


KAGE2 was recently released: Please use it and give feedback :)

- KAGE1 was released a couple of years ago and supported SNPs and indels
- KAGE2 is on bioRxiv and supports SVs
- KAGE works even better together with GLIMPSE
- **GPU-support** for insanely fast genotyping

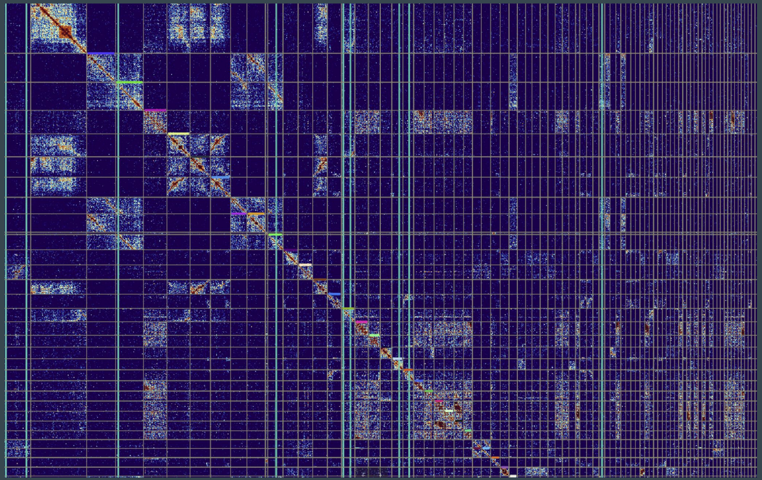
Why does speed matter?

- All of us Project: Genotype a million individuals
- Your project: Genotype a few hundred animals or plants?
- Also: Accuracy increase with pangenome size, current methods don’t scale



KAGE was built with BioNumPy

- Python-based, but >10x faster than PanGenie and other tools written in C
- Another tool we are building with BioNumPy is a **scaffolder**



If anyone is interested in scaffolding, please talk to me later

- Work by me, Knut Rand and Geir Kjetil Sandve
- KAGE is available at https://github.com/kage-genotyper/kage/
- Happy to answer questions :)