SevenBridges

BUILDING A BETTER REFERENCE

Conventionally, reference genomes are recorded as long sequences of letters representing DNA. At locations along the sequence where multiple variations occur in a population, only a single allele is selected for the reference. This simplifies the representation, but at the expense of throwing away diverse population information.

There is a better way:

build a pan-genome graph containing population information.

sevenbridges.com

FAMILY TRIO TRACING
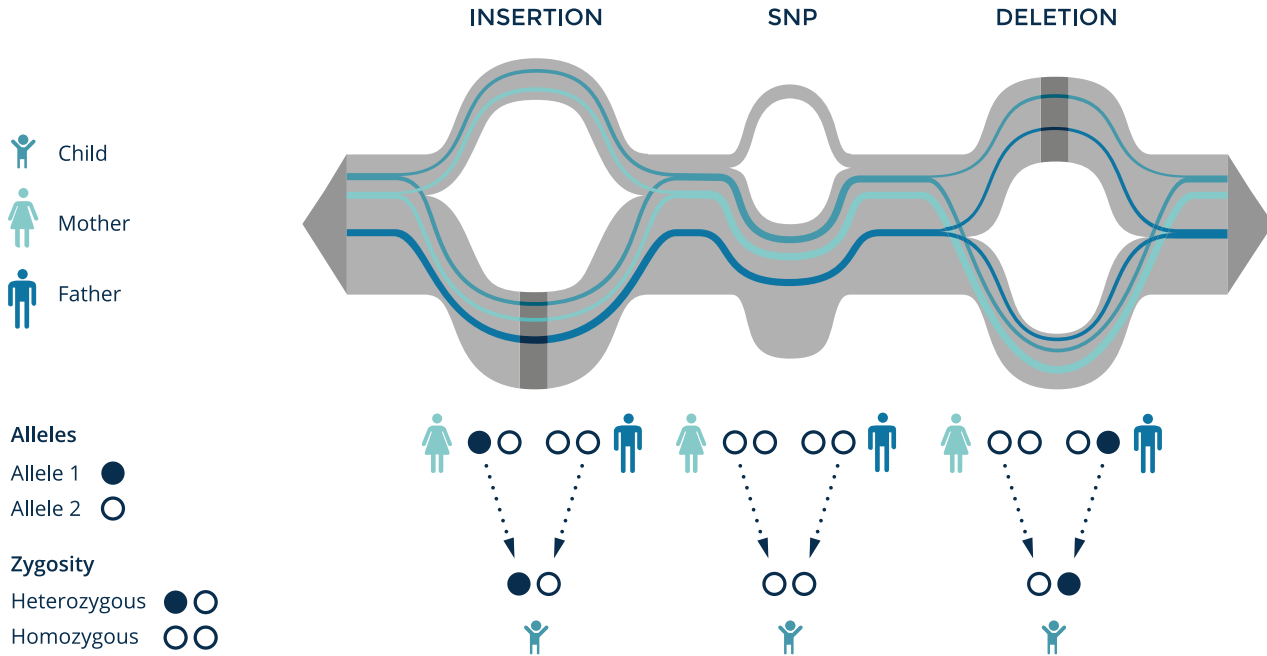
We can track the paths of individuals from a family trio on the graph, define the zygosity of their variants and inheritance of the variants in the child. By inspecting the much smaller set of locations where genomic read sequence data fails to match the graph, we can also quickly narrow down the novel mutations observed in the child.

By preserving population scale genomic variant information in a single data structure, the pan-genome graph provides an intuitive representation for genomic data and a powerful tool for new discovery.

Get in touch & find out more at
sevenbridges.com/graph

SevenBridges

sevenbridges.com
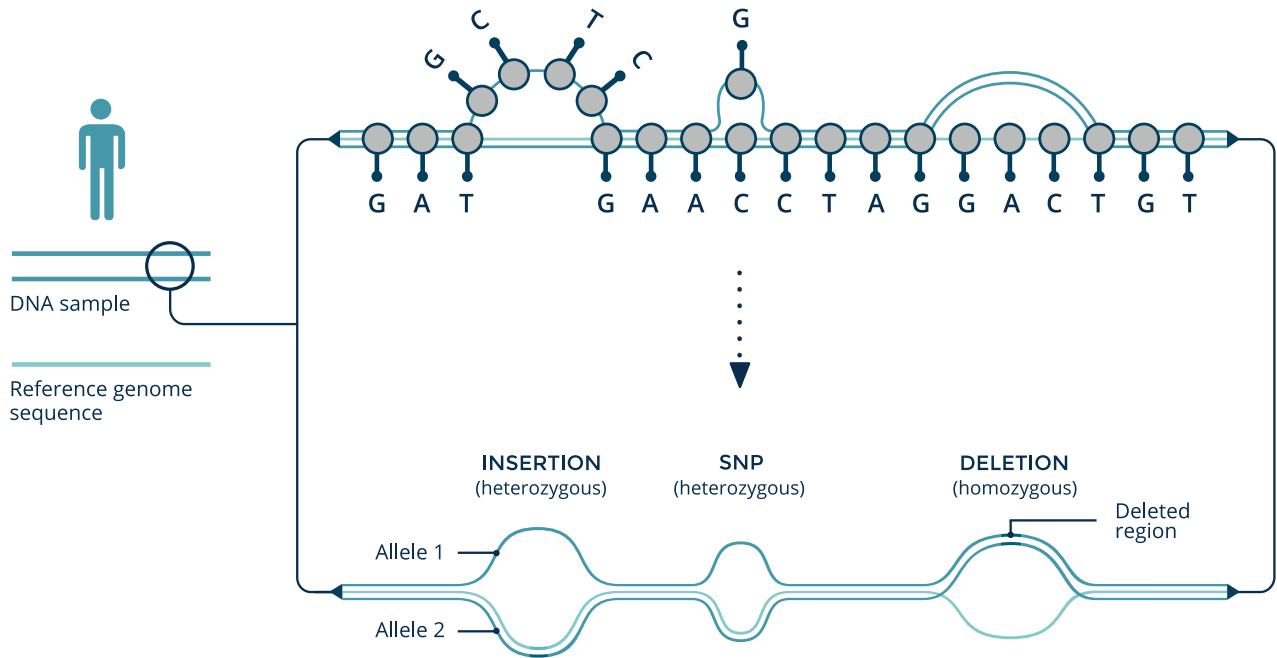
BUILDING A BETTER REFERENCE

Conventionally, reference genomes are recorded as long sequences of letters representing DNA. At locations along the sequence where multiple variations occur in a population, only a single allele is selected for the reference. This simplifies the representation, but at the expense of throwing away diverse population information.

There is a better way:

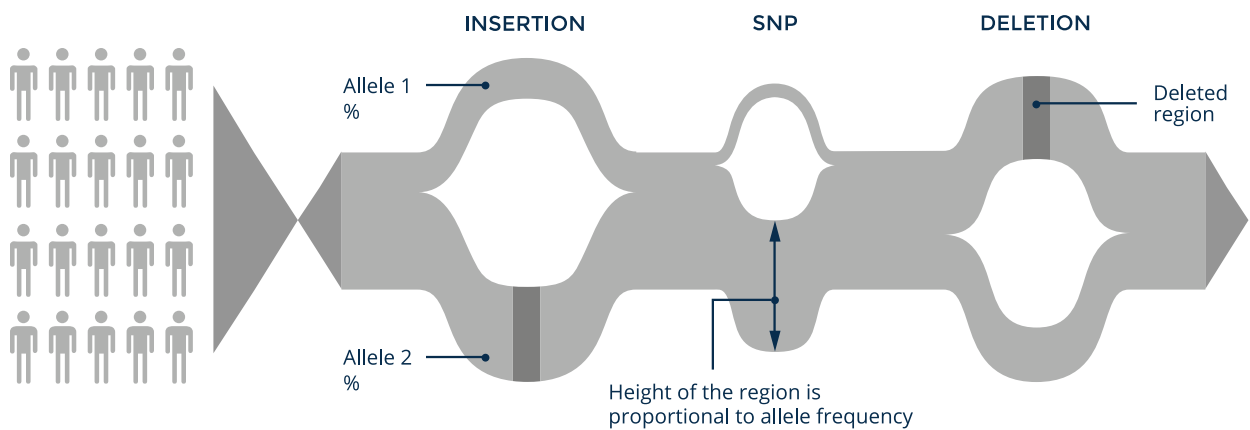Build a pan-genome graph containing population information.

FROM GENOMIC VARIANTS TO A GRAPH

INSERTION
(heterozygous)

SNP
(heterozygous)

DELETION
(homozygous)

Deleted region

Allele 1
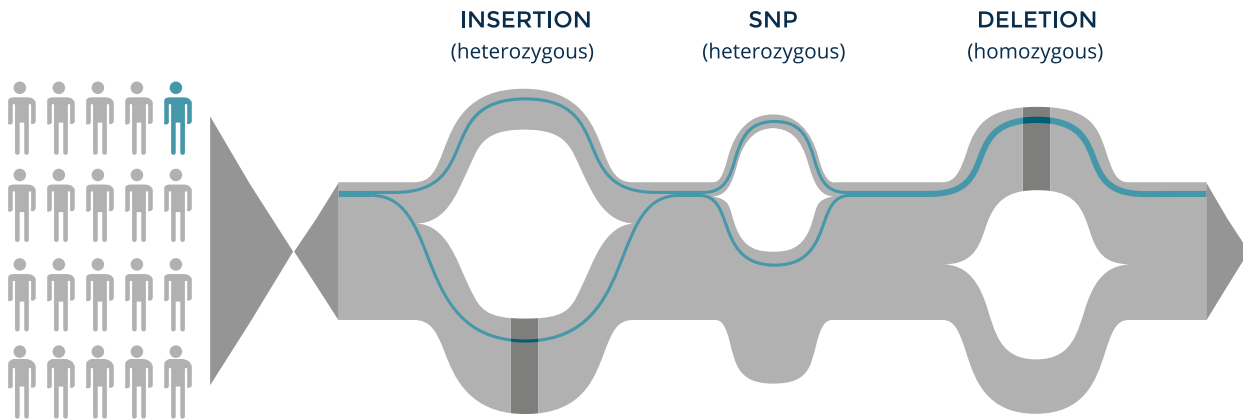
Allele 2

DNA sample

Reference genome sequence

We can build a genome graph by starting with a conventional genomic reference sequence, and adding new variants as additional branches, which differs from the reference sequence, but later rejoin it. The power of the graph structure is that it can be updated as new haplotypes are discovered.

WORLD POPULATION GRAPH

INSERTION

SNP

DELETION

Allele 1 %

Allele 2 %

Deleted region

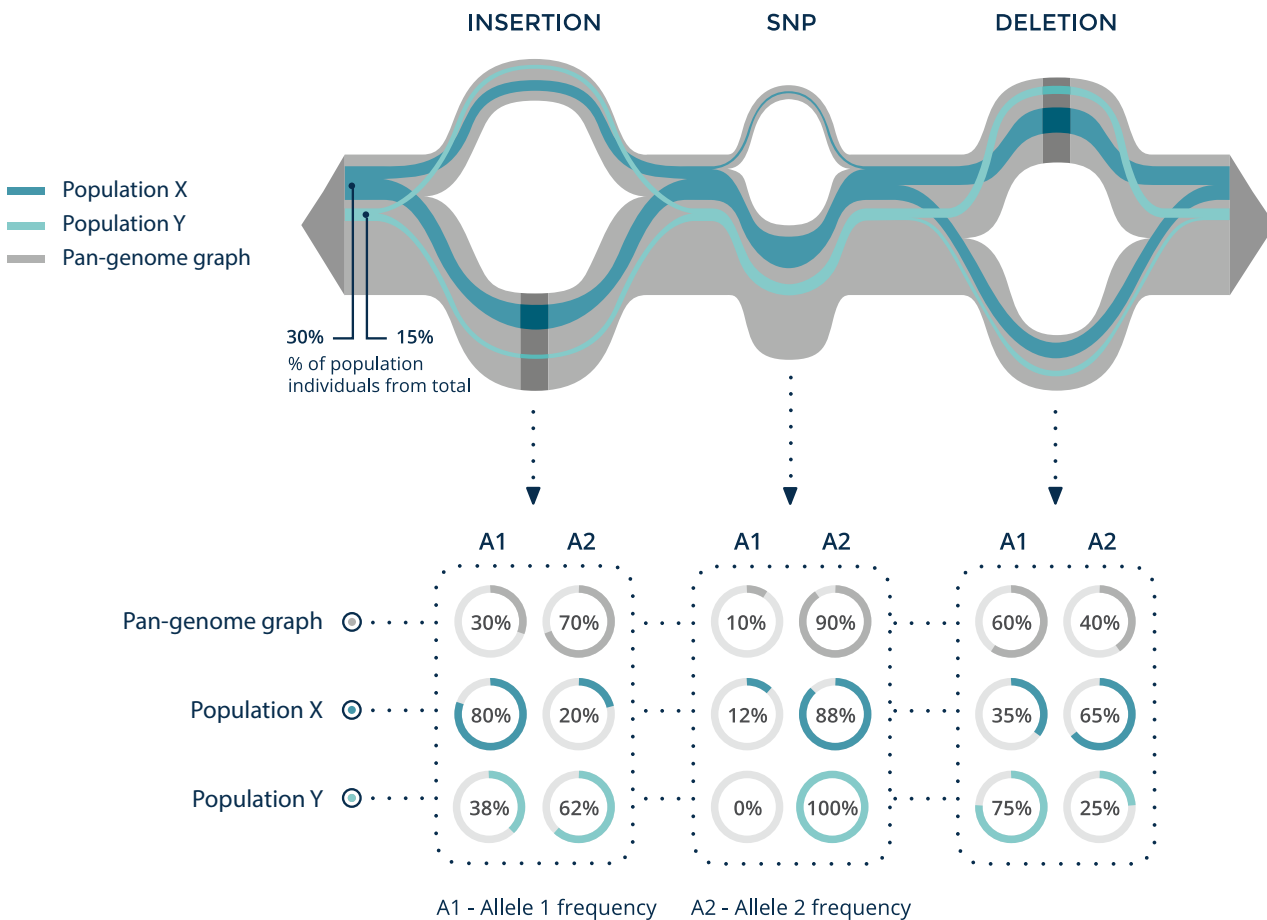Height of the region is proportional to allele frequency

If we take all the genomic information available worldwide and put it into a single graph structure, we can build a world population or pan-genome graph. Each branch has an allele frequency associated with it, which measures how often it appears in the individuals included in the graph.

INDIVIDUAL TRACING

| INSERTION (heterozygous) | SNP (heterozygous) | DELETION (homozygous) |

The genome of each individual is a set of paths along the population graph and can be tracked to compare his or her variants and their allele frequencies to the population frequencies. On the other hand, each individual can be anonymized by not storing path information, whereby his or her variants are aggregated to the population graph with no link to the individual.

POPULATION TRACING

INSERTION   SNP   DELETION

— Population X
— Population Y
— Pan-genome graph

30% — 15%
% of population
individuals from total

	A1	A2	A1	A2	A1	A2
Pan-genome graph	30%	70%	10%	90%	60%	40%
Population X	80%	20%	12%	88%	35%	65%
Population Y	38%	62%	0%	100%	75%	25%

A1 - Allele 1 frequency   A2 - Allele 2 frequency

Here one can see two populations, together with the allele frequencies of each of their variants. This type of analysis can be used to identify a rare SNP which is present in the Population X but absent in the Population Y.