The Human Genome Variation Map

Genomics at the scale of everyone



Adam Novak
Benedict Paten
David Haussler

Baskin Engineering | UC Santa Cruz

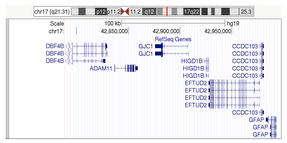
UNIVERSITY OF CALIFORNIA SANTA CRUZ | Genomics Institute

Making progress against the really tough diseases will require learning across millions or billions of genomic features, and consequently millions or billions of individual people's genomes.

The Human Genome Variation Map (HGVM) project aims to create a genomic reference that is representative of humanity, and a toolchain for analyzing people's genomes in the context of that reference.

https://hpsrepository.asu.edu/handle/10776/2130

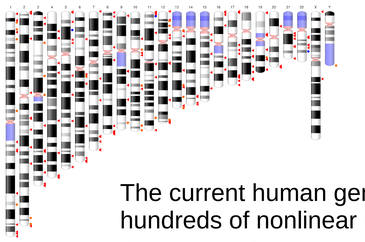
Thomas Hunt Morgan invented the linear genome in 1911.



http://genome.ucsc.edu/cgi-bin/hgTracks

Genomics software works in a linear space.

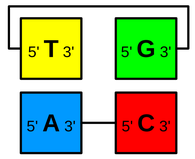
Linear References



The current human genome assembly contains hundreds of nonlinear regions (red).

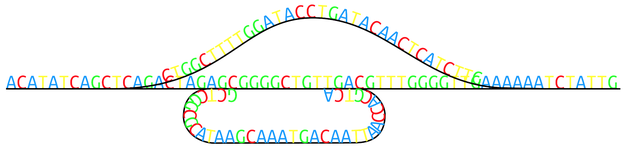
http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

Morgan, Thomas Hunt. "Random segregation versus coupling in Mendelian inheritance." Science (1911): 384-384.



A sequence graph is a set of positions, each with two sides. Adjacencies connect pairs of sides.

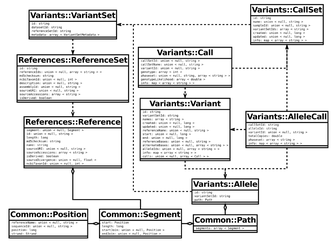
Graph References



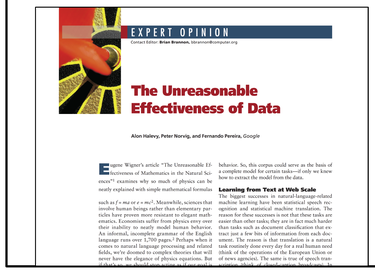
We can merge many genomes into a sequence graph and use it as a reference.

Global Alliance for Genomics & Health

The Global Alliance for Genomics and Health is designing a graph-reference-capable API to share genomic data.



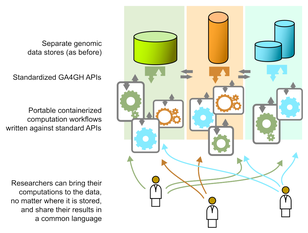
GA4GH and Data Sharing



Sharing data leads to larger data sets and more powerful analyses.

Halevy, Alon, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data." Intelligent Systems, IEEE 24, no. 2 (2009): 8-12.

Paten, Benedict et al. "The NIH BD2K Center for Big Data in Translational Genomics (CBDTG)." Journal of the American Medical Informatics Association (2015): Submitted.



The HGVM project will let bioinformatics analyses move to the data.

Distributed Analysis

This will allow the statistical power of arbitrarily large numbers of samples to be directed against hard genomics problems.

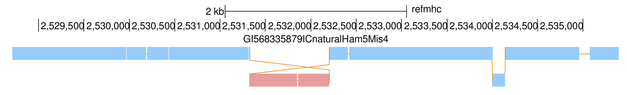
http://upload.wikimedia.org/wikipedia/commons/2/22/PS20andPS10.jpg

GTGACCGATCG**C**TACGTGCTACGGACT
CCGATCGC**TACGTGCTACG**

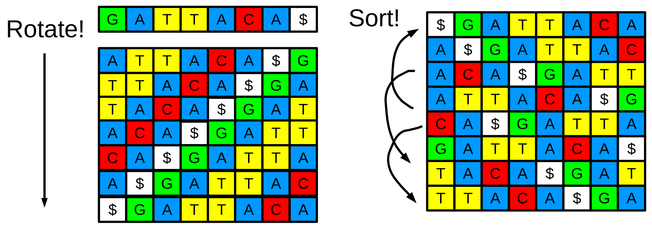
Individual bases in the HGVM graph can be uniquely identified by their local contexts.

Context-driven Mapping

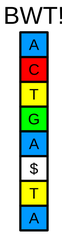
This makes the detection and description of structural rearrangements in people's genomes much easier.



Ferragina, Paolo, and Giovanni Manzini. "Opportunistic data structures with applications." In Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on, pp. 390-398. IEEE, 2000.



Compressed Self-Indexes



Burrows-Wheeler Transform: Characters grouped by context.

Can search for n-character substring in O(n) time.