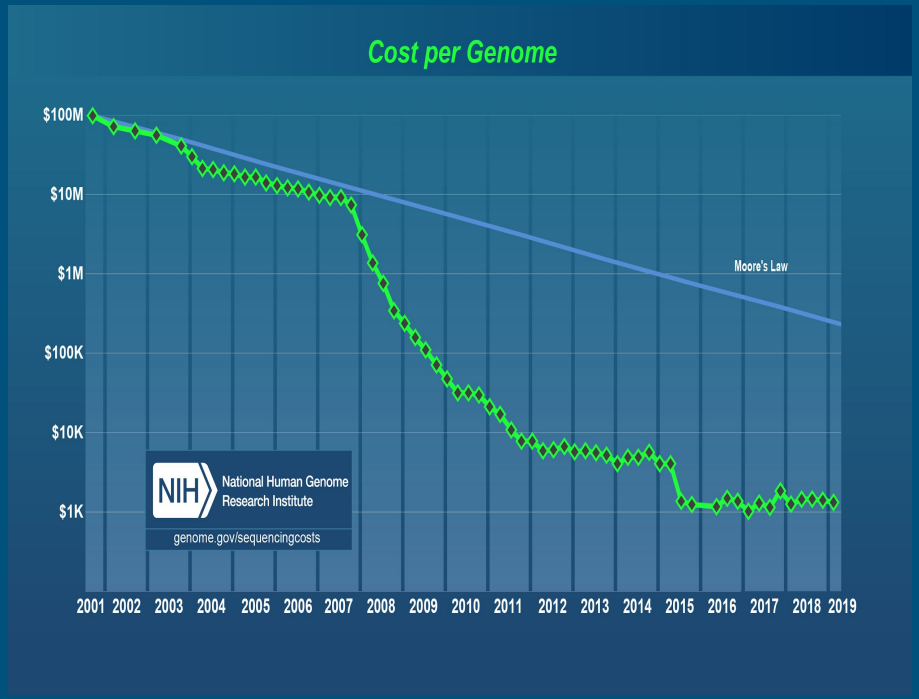
Resource-frugal analysis of whole genomes

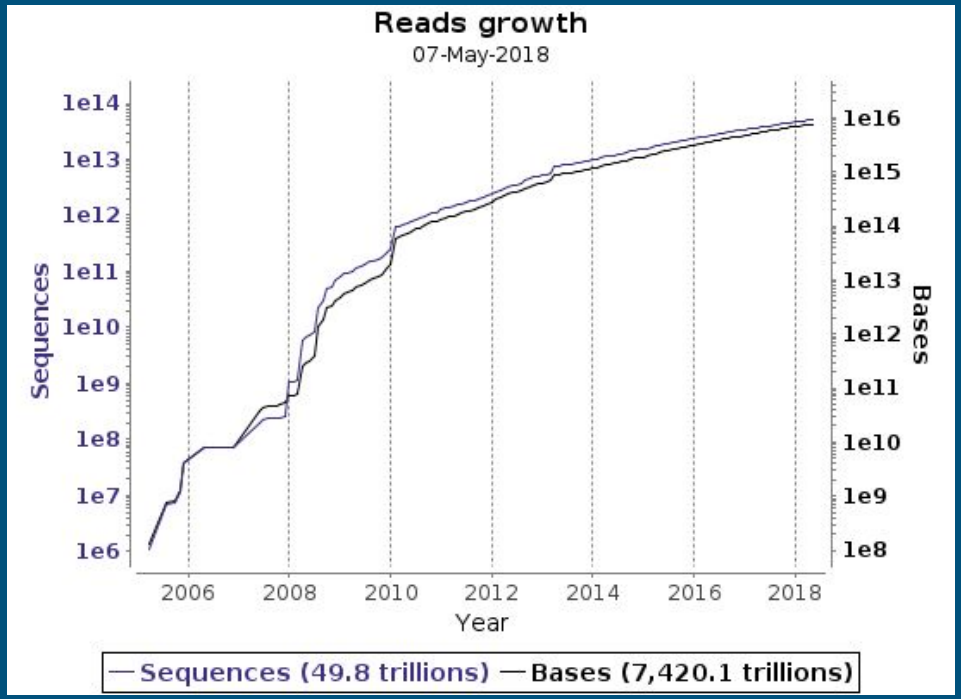
vs.

Sven Rahmann
Genominformatik, Universität Duisburg-Essen
& Informatik XI, TU Dortmund

Cost and numbers of sequenced genomes



Source: NIH



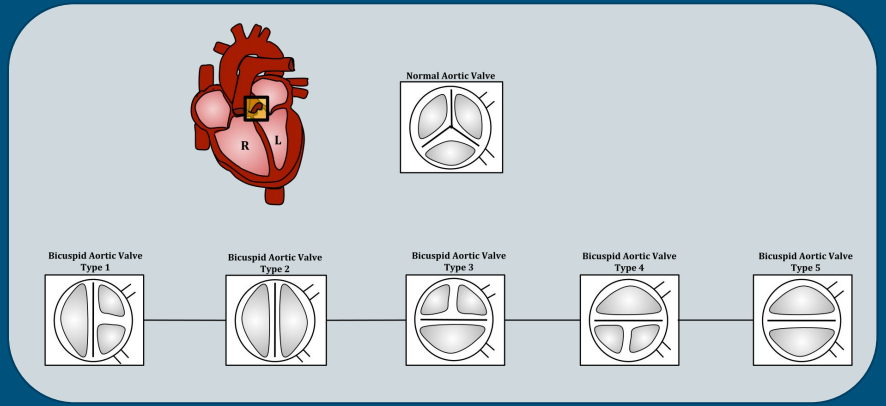
Growth of Sequence Read Archive (SRA),
Source: European Bioinformatics Institute (EBI)

Bicuspid aortic valve (BAV): a heart condition

- 250 BAV genomes
- 250 controls
- 100 Gbp per sample (30x coverage)
- 50 TB of data in total

Questions:

- Genetic features associated to BAV?
- Genetic features associated to particular subtypes of BAV?

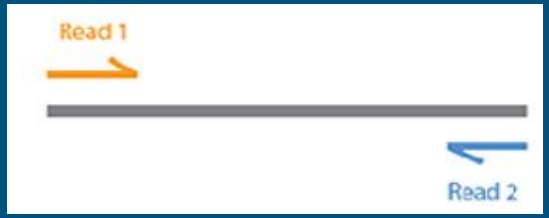


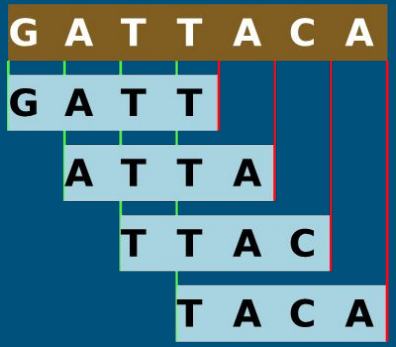
Source: https://en.wikipedia.org/wiki/Bicuspid_aortic_valve#/media/File:Bicuspid_Aortic_Valve.svg

(Collaboration with University Hospital Hamburg Eppendorf)

Approach: k-mer counting

- Genome is obtained as "short paired-end reads" (2x150 bp from 500 bp - 1000 bp DNA fragments)
- Classical approach:
  - Origin of each read is located on genome.
  - Information at each genome position (3.1 Gbp) is summarized and evaluated ("variant calling").
- Our approach: "alignment-free" or "*k*-mer based":
  - Partition reads further into *k*-mers, count them in each sample.
  - Select *k*-mers where the count is low in one class and high in the other class
  - Challenge of scale: 10 billion *k*-mers x 500 samples; count table does not fit in memory.





Desired results

- Small matrix of *k*-mers (↦ genes)
  with low presence in one class
  and high presence in the other class

- Assemble *k*-mers to obtain
  longer genomic sequences;
  biological interpretation

- Solve computational & statistical issues:
  - full table never instantiated
  - extremely-multiple testing
  - which test(s) ?