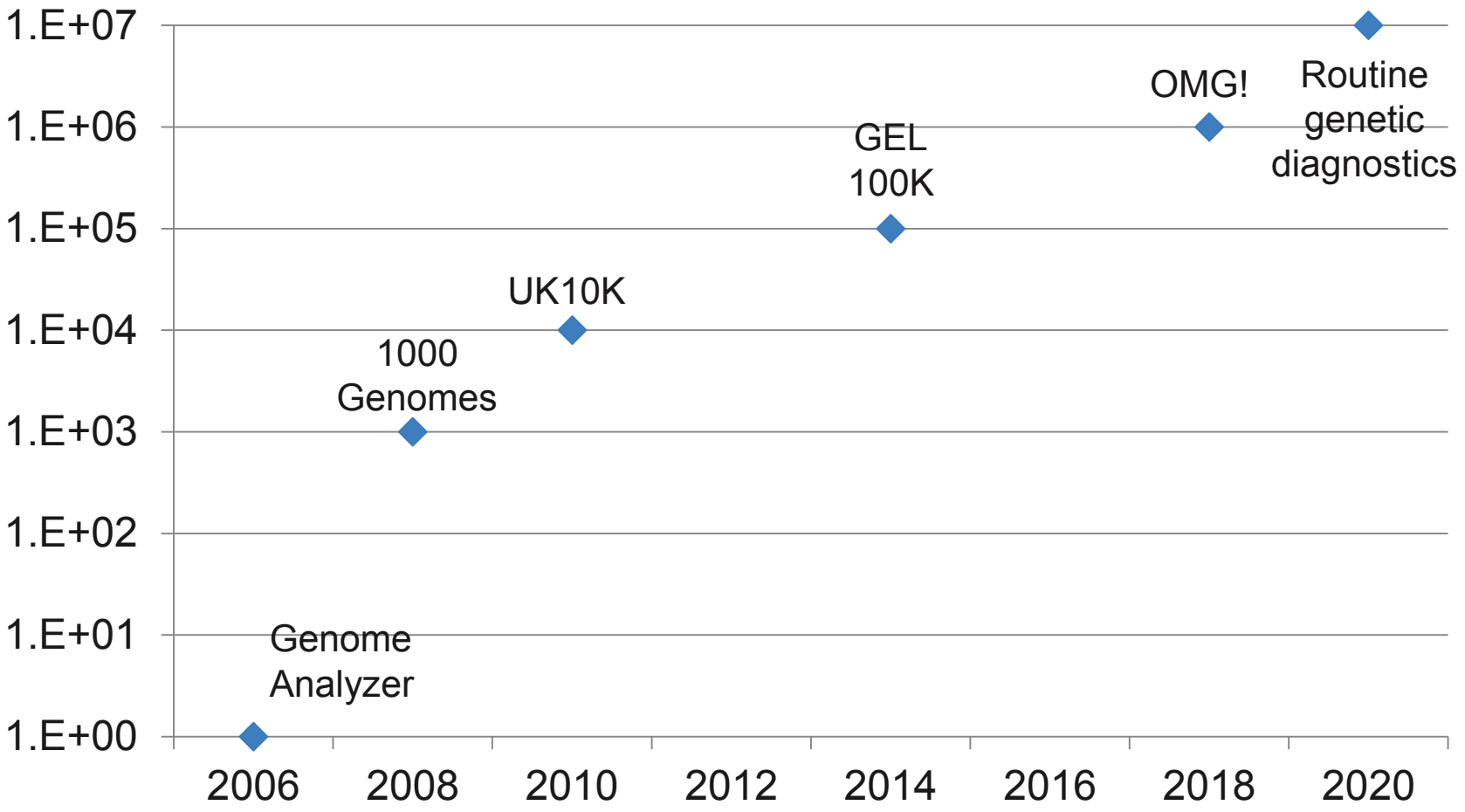
Compression

Come Raczy
Illumina, Inc

Evolution of human genome project sizes



illumina

What make a good compression format?

- It depends on the use cases
  - The obvious data footprint is not necessarily directly relevant
  - Other factors like time, money or risks are often critical
  - Broad acceptance of a single format might be the most important criteria
- Use case: data transfer from instrument to data processing facility
  - Data footprint defines the minimum network bandwidth requirements
  - Suboptimal compression per cycle enable shorter Turnaround times (TAT)
  - Resilience to data loss and data corruption is critical
- Use case: short term storage for data analysis
  - Trade off between IO reduction and CPU usage increase for (de)compression
- Use case: medium term storage for data queries
  - Granularity of decompression and accessibility of data of interest
- Use case: long term storage for archival
  - The overall cost is dominated by data footprint only

What data needs compression (made up numbers!)

		2016	2018	2020	2022
Factory Max/year	30X genomes	50K	150K	500K	1500K
	BAM volume	3PB	9PB	30PB	90PB
Instrument Max/year	30X genomes	2K	10K	50K	200K
	BAM volume	120TB	600TB	3PB	12PB
Cost of sequencing / genome		$500	$200	$75	$25
Cost of storing (genome-year)		$22/$9/$5	$16/$7/$4	$12/$6/$3	$9/$5/$2
Read length		2x150bp	2x200bp	2000bp	10Kbp
Analysis		remote	mixed	local	local
TAT		2h	30mn	0	0
Organisms mixture		Single	Few	Many	Many

Lossy or lossless?

- Base calls and quality scores are a lossy image compression scheme
- The community regularly decides to discard irrelevant information (e.g. non PF, 2nd base call, 60 quality bins, etc.)
  - Today's consensus: 2 to 8 quality bins, base calls for reads PF
  - Quality scores could be Boolean or discarded altogether
  - Detailed base calls could be provided only for sequences of interest
  - Aggregate information (e.g. k-mer counts) could replace some base calls
- Parallel with JPEG: if you can't see it, the loss is acceptable, or required
  - The "eyes" are secondary analysis software and interpretation processes

Quality score compression

- Broad variation in quality score distributions across platforms
  - Illumina has a strong focus on data quality (%Q30 >> 70%)
  - Other technologies trade other strengths (e.g. read length) for lower quality
- Radically different properties of quality scores across technologies
  - SBS: mostly monotone decreasing along the read
  - Nanopores: sequence dependent
- For a given technology, broad variability in compressibility
  - Variability from the hardware (e.g. detector sensitivity, ambient noise)
  - Variability from the recipe (e.g. noise inherent to higher throughput)
  - Variability from the consumables (e.g. reagents, flowcells, membranes)
- Lossy strategies depend on the use-case context
  - For unaligned data: mostly limited to the number of bins
  - For aligned data: possibility to aggregate by reference position

Compressing Illumina q-scores (4 bins)

- Uncompressed would be 2 bits/q-score
- RLE + gzip: 0.5 to 1.0 bits/q-score (2-4x compression)
- 1st order Markov model: 0.3 to 0.7 bits/q-score (3-7x compression)
  - Many methods using the quality from the previous cycle lead to very similar compression ratios
  - 1st order arithmetic encoding is very close to the actual entropy
  - Experimentation with higher order models doesn't seem to lead to any significant improvement
- Anecdotal evidence indicates that aggregates per reference position doesn't affect variant calling
- With good quality data, reduction to 3 or even 2 quality bins has limited effect on variant calling
- With really good data, read filtering would almost guaranty good-enough base q-scores – making the base q-score unnecessary (e.g. replacing all base q-scores with the usable length in the read)

Base call compression

- As with q-score compression, data quality affects compressibility
  - Even with a high proportion of >Q30 data, the small proportion of lower quality bases can lead to a large number of uncompressible errors
- Coverage is the main driver for compressibility
  - Broad range of coverage from less than 1 to more than 10K
  - Vast majority of datasets in the range 20-80
  - Data from several samples from the same species usually compresses like a single sample at the cumulative coverage
- Genetic diversity of the input samples is the other important factor
  - Human DNA only is easy to compress
  - Microbiomes can lead to a lot of diversity
- Lossy compression
  - If a base is interpreted as an error, it probably doesn't matter
  - In many regions, as soon as there is enough data to support a call, it doesn't really matter what the exact coverage is
  - In many regions we know that only rare and very specific conditions will

Compression of Illumina base calls (2x150bp short insert)

- Methods based on alignment
  - Aggregated information based on CRAM
  - Various publications related to compression of DNA sequences
- Burrows-Wheeler Transforms + next base prediction
  - BEETL: in-house tool to incrementally build the BWT of the reads https://github.com/BEETL/BEETL
  - Effective reference-free compression: 0.4-0.5 bits/base (4-5x compression)
  - Very sensitive to the coverage
  - Works well across samples of the same species
- K-mer counting + next base prediction
  - Similar results to BEETL with appropriate value of K (24 for human)
  - Very stable between runs – with same species, same recipe, same platform

Q&A