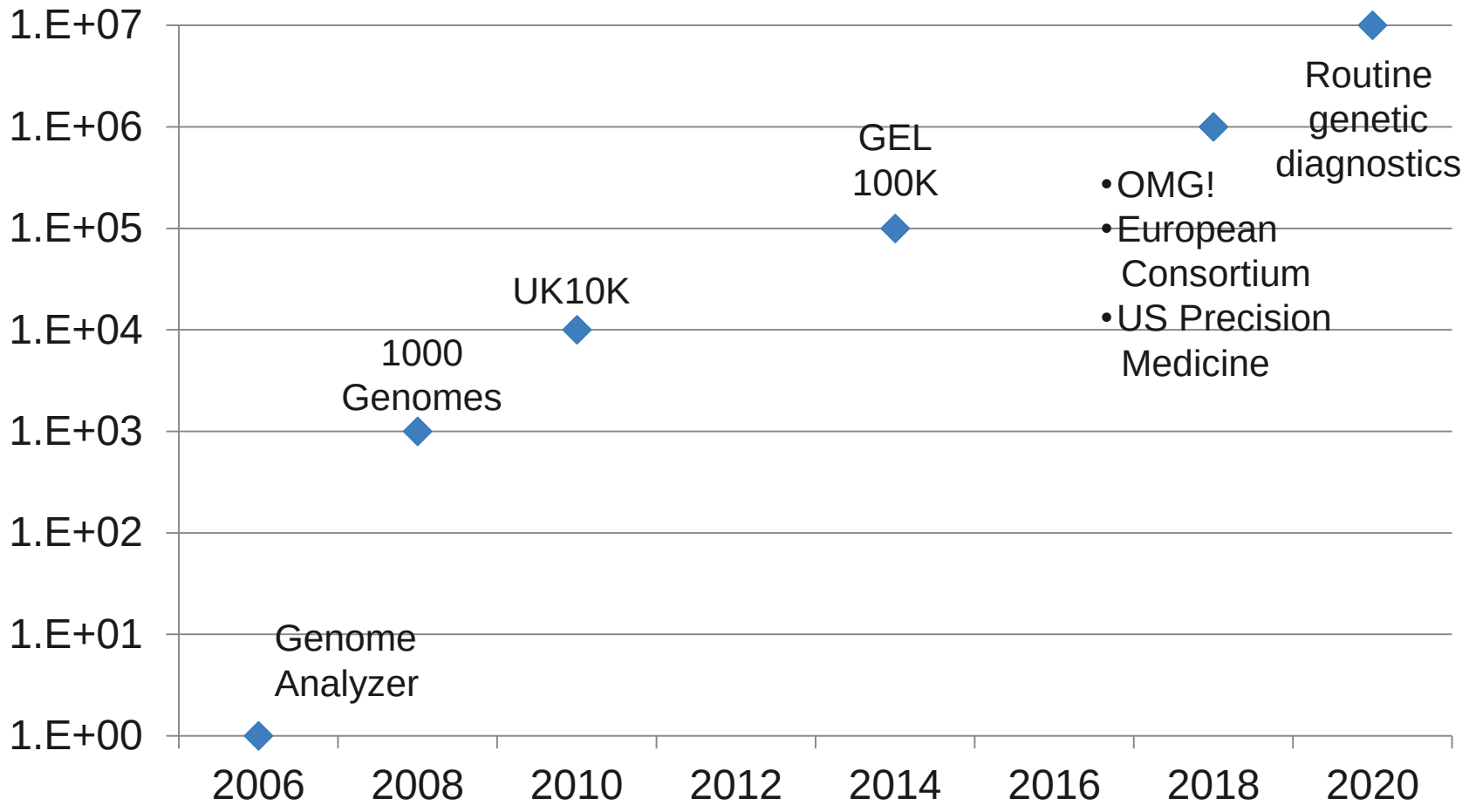
The role of compression in the genomics data life cycle

Come Raczy
Illumina, Inc

illumina®

Evolution of human genome project sizes



illumina®

What data needs compression (made up numbers!)

		2016	2018	2020	202x
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		$1000	$500	$200	$100
Cost of storing (genome-year)		$22/$9/$5	$13/$6/$3	$12/$6/$3	$10/$5/$2
Read length		2x150bp	2x150bp	2x250bp	1Kbp
Analysis		remote	mixed	mixed	local
Analysis TAT		2h	2h	30mn	0
Organisms mixture		Single	Few	Many	Many

Evolution of storage strategies

- 2008: whole set of TIFF images (>>100 bytes per base)
- 2010: channel intensities of extracted features (~20 bytes per base)
- 2012: unaligned reads with full base quality score (~1 byte per base)
- 2014: aligned reads with full metadata (BAM format) (6 bits per base)
- 2016: aligned reads with reduced base quality scores (4 bits per base)
- 2018: aligned reads with reduced metadata (CRAM) (3 bits/base)
- 2020: advanced read compression and aggregated base quality scores
- …
- 202x: Variant calls

Typical data life cycles

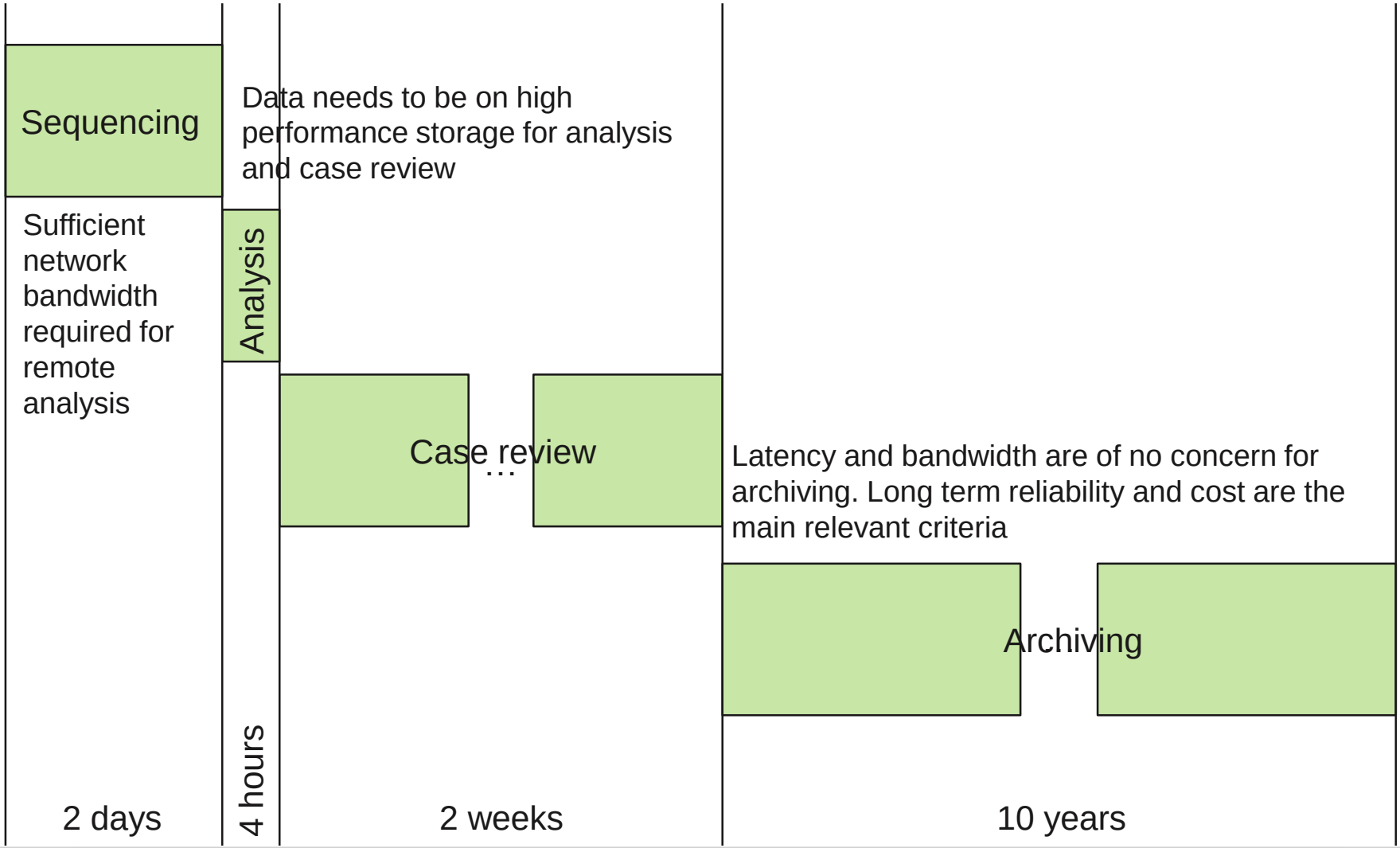
Capacity limited by available bandwidth

- Flexible on-demand usage
- Increased storage cost

Capacity limited by available hardware

- Optimal cost for steady load
- Increased organizational burden

Typical data life cycles



illumina®

Data size limitations
assuming 4 bits/base or 50GB/genome

- Maximum capacity of a 1Gbps network connection
  - 216 genomes/day
  - 78,840 genomes/year
- Minimal cost of storage on AWS
  - $12.60/genome/year on S3
  - $6/genome/year on S3 (one zone infrequent access)
  - $2.40/genome/year on Glacier
- Local storage solutions for 200 genomes/day
  - Live storage for 2 week: 150TB
  - Long term storage for 10 years: 40PB

Relative costs of storage

- In theory, these costs are << 5% of the overall costs assuming that
  - The bulk of the data gets quickly archived at low cost
  - Redundancy is very well controlled
  - Storage solutions are cost-effective
- For many users, actual storage costs are much higher
  - Sub-optimal retention policies and archival strategy
  - Redundant storage driven by the diversity of analysis workflows
  - Frequent use of over-priced storage solutions
- For most practical uses, storage costs should be much lower (<< 1%)
  - Adoption of more compact data formats (e.g. CRAM instead of BAM)
  - Introduction of lossy compression formats (e.g. lossy CRAM)
  - Deletion of all irrelevant read data (e.g. all reads not supporting selected variant calls)

Relative costs of storage
Example (rough estimates in a perfect world)

Operation	$ Abs.	$ Rel.	Comments
Sequencing	$500	63.3%	• Assuming large capacity factory • Varies with instrument and volume • Opportunities for major cost reductions
Case Review	$200	25.3%	• Up to 4 hours hands-on expert work • Varies with applications • Opportunities for major cost reductions
Sample Management	$50	6.3%	• Sample collection, transport, storage… • Impacts all other operations • Opportunities for operational optimizations
Data Storage	$30	3.8%	• Assuming 10 years archiving • Usually 2-4 time this cost • Optimal cost proportional to data footprint
Data Analysis	$10	1.3%	• All secondary and tertiary analysis • Based on AWS on-demand pricing • Opportunities to optimize workflows

Factors contributing to spurious storage costs

- Fast evolution of technology
  - Sequencing and analysis pushing the value towards variant calls only
  - Storage costs still decreasing steadily
  - Explosion of quickly evolving file formats and compression solutions
- Uncertainty regarding projected needs for data retention
  - Balancing regulatory needs with practical usefulness
  - Small extra cost on storage is a safe way to mitigate risks
- Cost of engineering change
  - Many workflows consume data in multiple different formats
  - Unexpected consequences for modifying a data format or a workflow
- Prioritization of infrastructure and operational development
  - Data management and data retention policy are low priority
  - Missing features in other areas often have a much higher impact on value
  - Producing 50K samples per year presents major challenges

Examples of existing file formats and compression solutions

- Unaligned reads
  - Gzip/bzip2 compressed FASTQ: 7 bpbp (~90GB per 30x human genome)
  - Instrument native base call formats: 3 bpbp
  - PetaSuite: proprietary transparent sort-based FASTQ compression (3 bpbp)
  - Multitude of open source reference-based and BWT compression tools
- Aligned reads
  - BAM: 5-6 bpbp
  - CRAM lossless: 3-5 bpbp
  - CRAM lossy: 2-3 bpbp
  - Occasional introduction of new formats and layouts (e.g. column-based)
- Variants
  - VCF (rapidly evolving): 200-1000MB per human genome (single sample)
  - BCF (rapidly evolving): 100-500MB per human genome (single sample)
  - Several emerging solutions for multi- and many-samples variant stores

Q&A