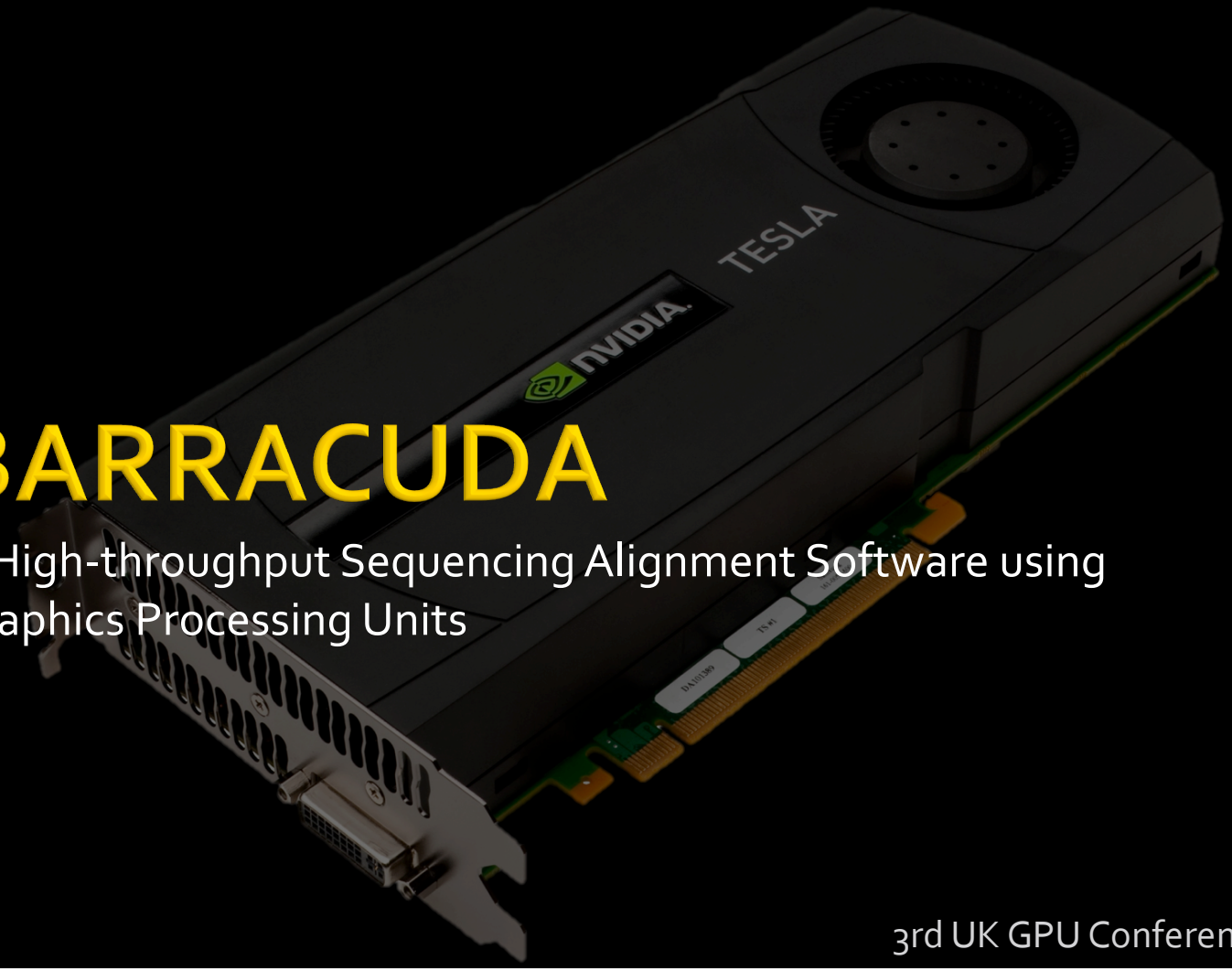


BARRACUDA

A High-throughput Sequencing Alignment Software using Graphics Processing Units

Brian Lam

3rd UK GPU Conference, 14th Dec 2011

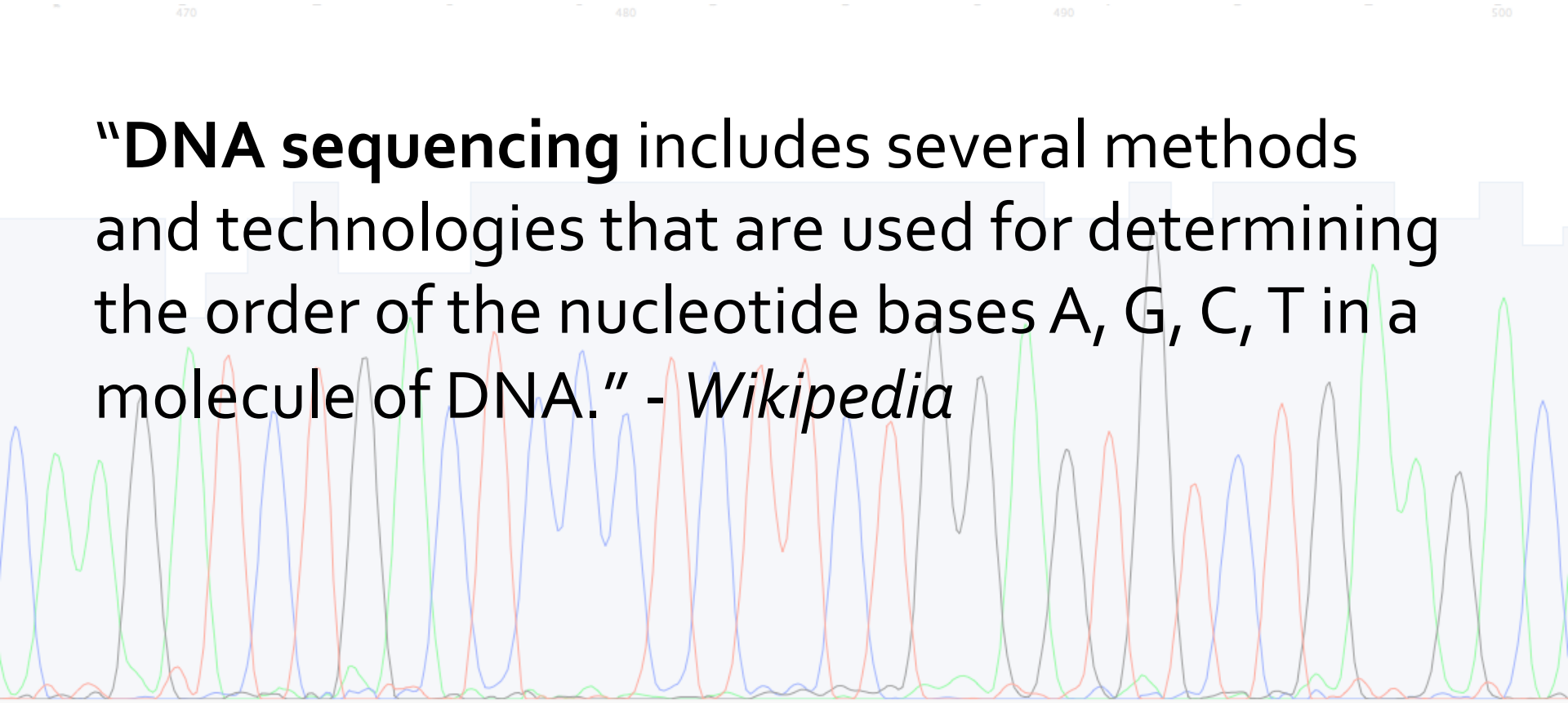
UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

The Genome and DNA

- A genome of a living organism is a complete repertoire of genetic programs (including genes and other regulatory regions) that encode all its functions and development
- Individual functions are determined by genes, which are coded by strings of deoxyribonucleic acids (DNA), comprising of adenine (A), guanine (G), cytosine (C), and thymine (T), joined together in a specific order (e.g. ACCATG)
- A Human genome has 3 billion DNA nucleotide bases (or bp)

DNA sequencing



"**DNA sequencing** includes several methods and technologies that are used for determining the order of the nucleotide bases A, G, C, T in a molecule of DNA." - *Wikipedia*

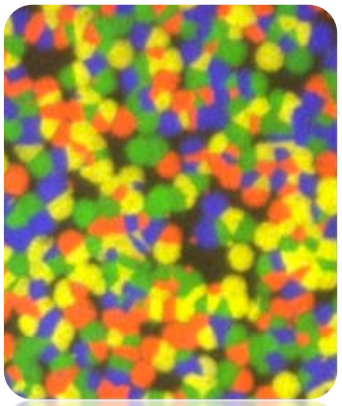
High-throughput DNA sequencing

- Commercialised in 2005 by 454 Life Science
- Sequencing in a massively parallel fashion

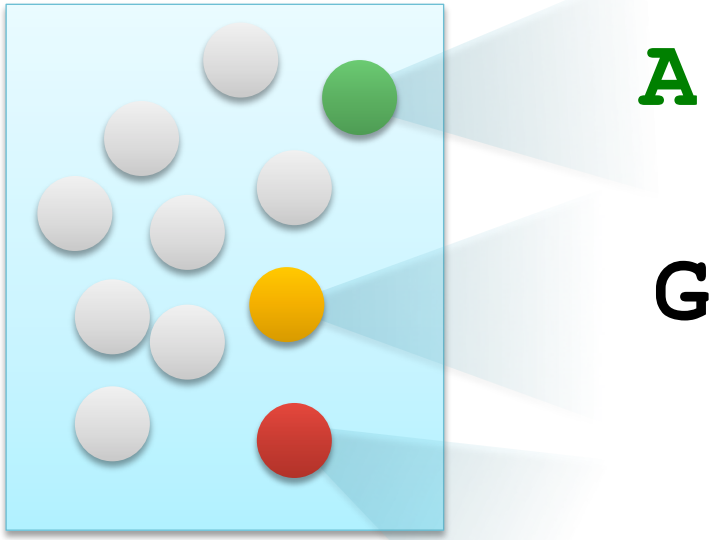
How it works

- An example: Whole genome sequencing
  - Extract genomic DNA from blood/mouth swabs
  - Break into small DNA fragments of 200-400 bp
  - Attach DNA fragments to a surface (flow cells/slides/ microtitre plates) at a high density
  - Perform concurrent "cyclic sequencing reaction" to obtain the sequence of each of the attached DNA fragments

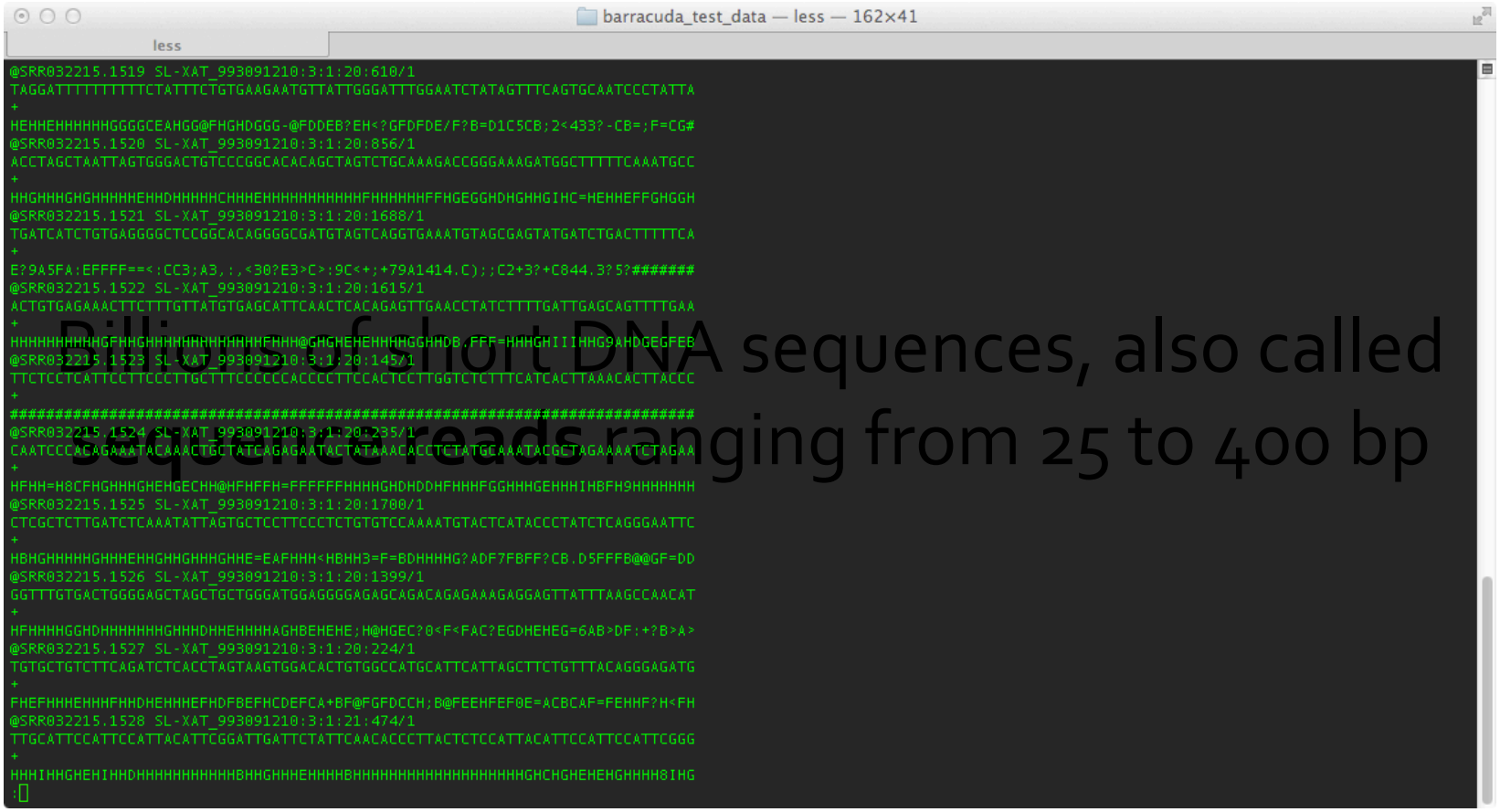
An Illumina HiSeq 2000 can interrogate 825K spots / mm²



Capturing the sequencing signals



What do we get from the sequencers

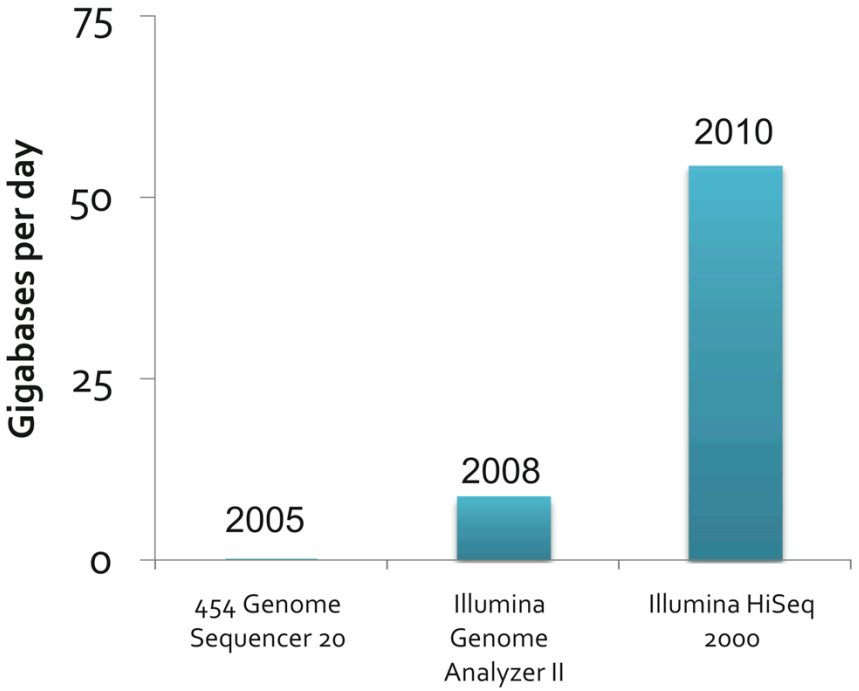


Billions of short DNA sequences, also called sequence reads ranging from 25 to 400 bp

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

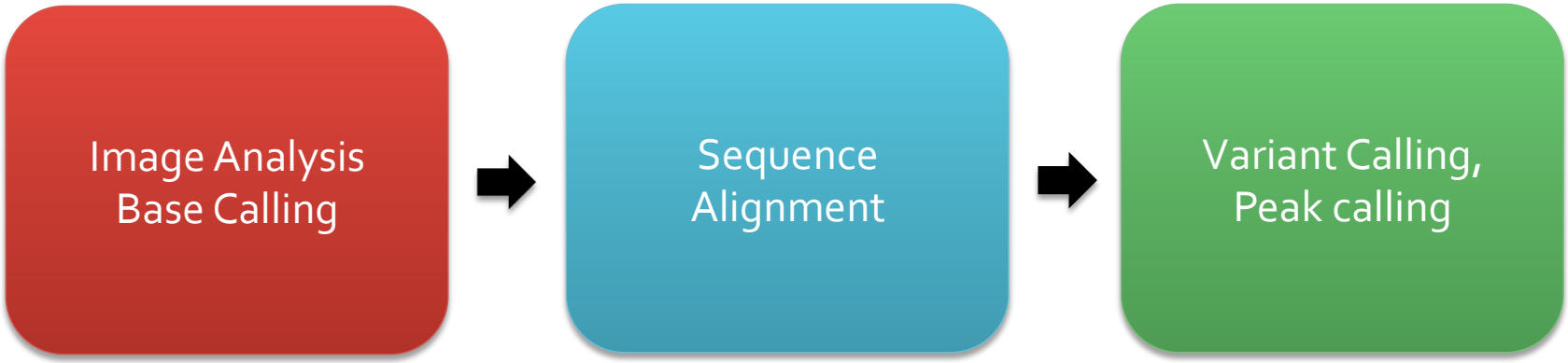
The throughput of HTS has increased dramatically



UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

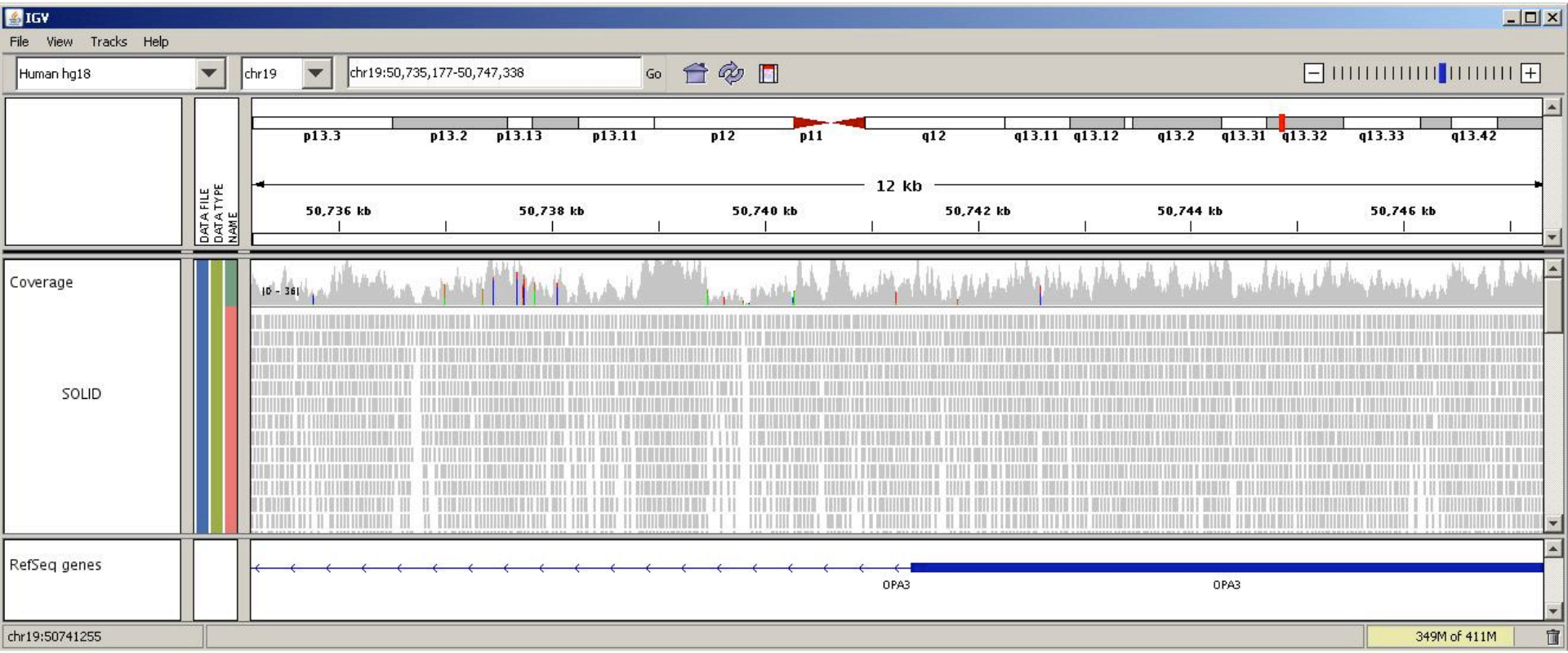
Current bioinformatics pipeline



UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Sequence alignment



UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Many-core computing

- Parallel computing using processors that contain a 'large number' on processing units on a physical die
  - Examples
    - NVIDIA Tesla M2090 - 512 CUDA cores (GPGPU)
    - AMD Radeon HD 7900 series - 2048 cores? (GPGPU)
    - Intel Knights Corner co-processor – 50x+ **x86** cores

Why bother?

- Low capital cost and energy efficient
  - Dell 12-core workstation (144 GFLOP/s): £5,000, ~1kW
  - Dell 40-core computing cluster (480 GFLOP/s): £ 20,000+, ~6kW
  - NVIDIA Tesla C2070 (500G FLOP/s): £1,500, ~0.2kW
  - NVIDIA Geforce GTX 590 (1 TFLOP/s): £400, ~0.4kW
- Many supercomputers now also contain multiple GPU nodes for parallel computations

GPU in bioinformatics

- Examples:
  - CUDASW++ → 6.3X
  - MUMmerGPU → 3.5X
  - GPU-HMMer → 60-100x

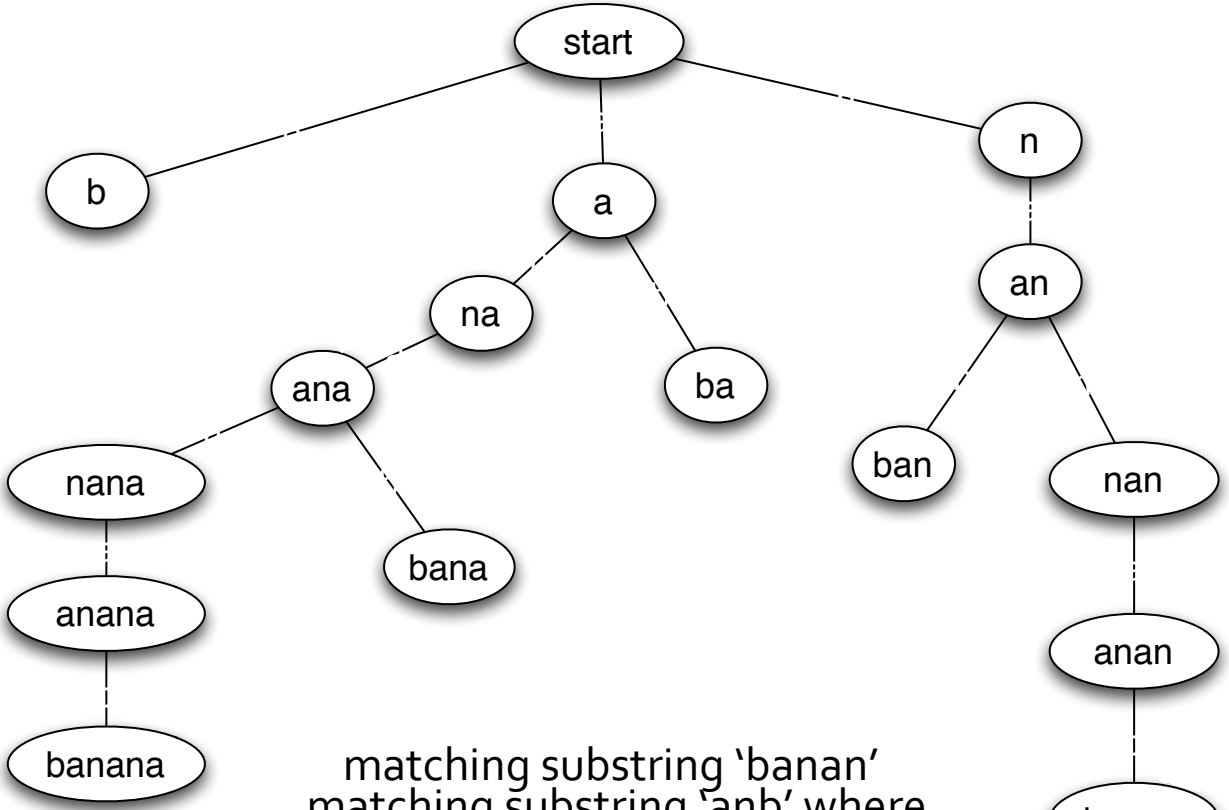
The BarraCUDA Project

- The main objective of the BarraCUDA project is to develop a software that runs on many-core architectures
  - i.e. to map sequence reads the same way as they come out from the HTS instrument

Burrows-Wheeler transform

- Originally intended for data compression, performs reversible transformation of a string
- In 2000, Ferragina and Manzini introduced BWT-based index data structure (FM-index) for fast substring matching at $O(n)$
- Sub-string matching is performed in a tree traversal-like manner
- Used in major sequencing read mapping programs e.g. BWA, Bowtie, Soap2

How it works – a backward search algorithm

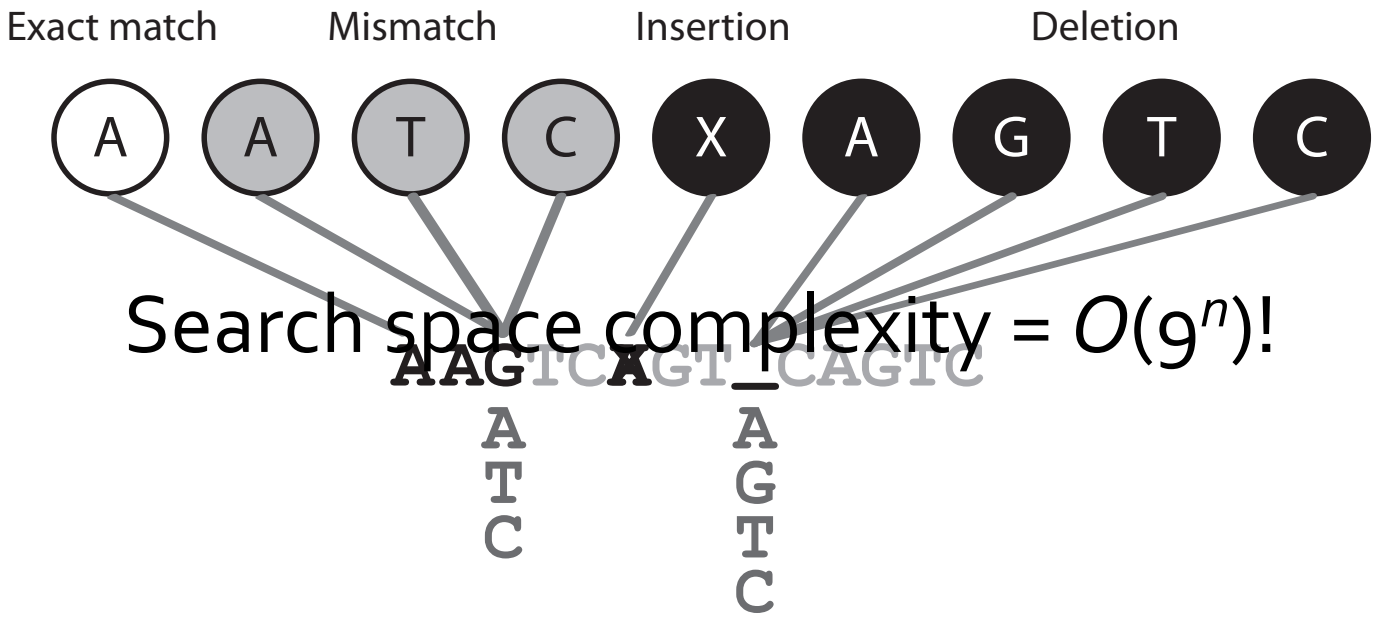


matching substring ‘banan’
matching substring ‘anb’ where
‘b’ is subsituted with an ‘a’

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Inexact matching requires base substitution within the query substring



Search space complexity = $O(9^n)$!

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

BarraCUDA started its life as a GPU version of BWA

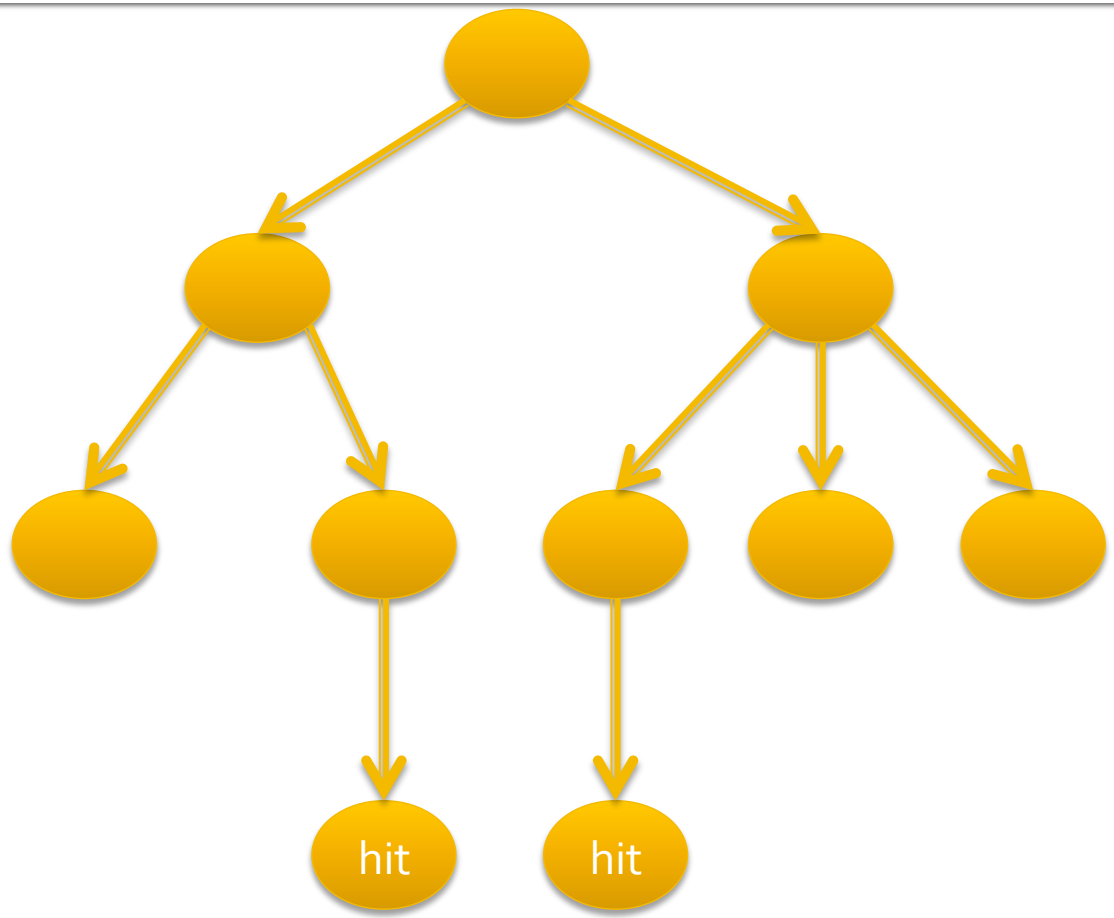
- Ported BWA to CUDA with simple data parallelism
- Used mainly the GPU for sequence mapping

- And it _partially_ worked!
  - 10% faster than 8Cs @ 3GHz
  - BWA uses a greedy breadth-first search approach (takes up to 40MB per thread)
  - Not enough workspace for thousands of concurrent kernel threads (@ 4KB) – i.e. reduced accuracy

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

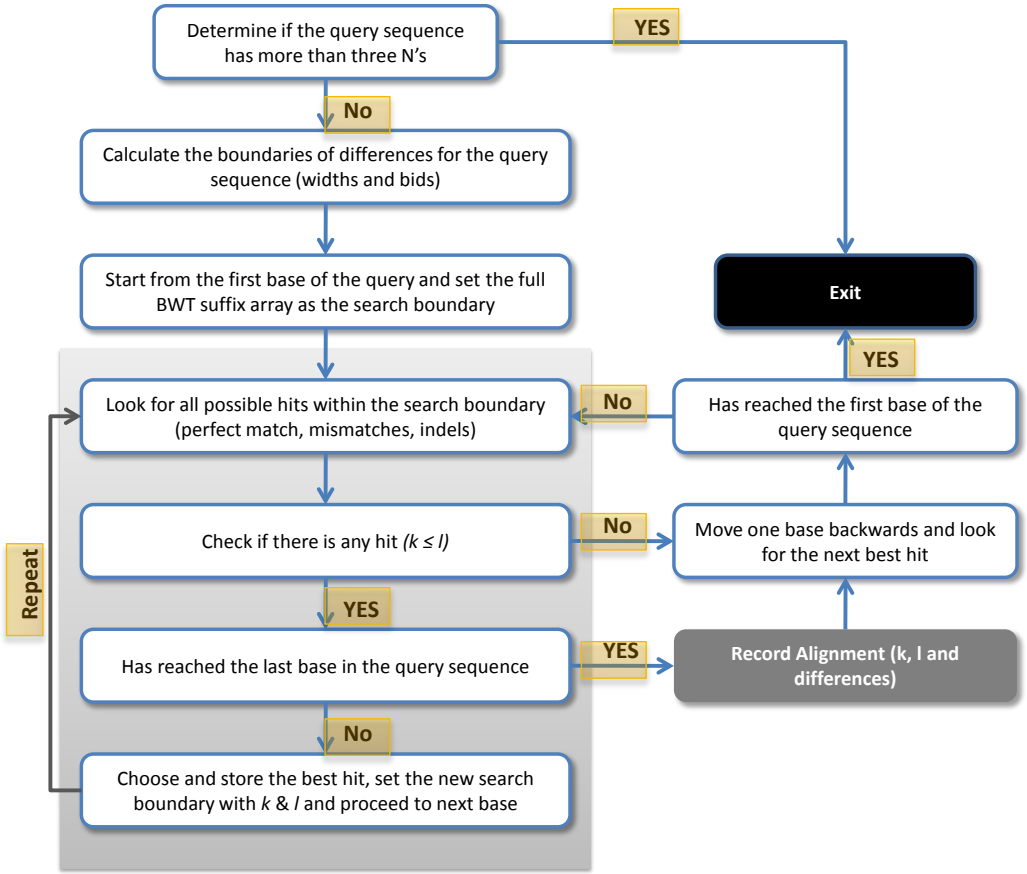
BarraCUDA uses a depth-first search approach



UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

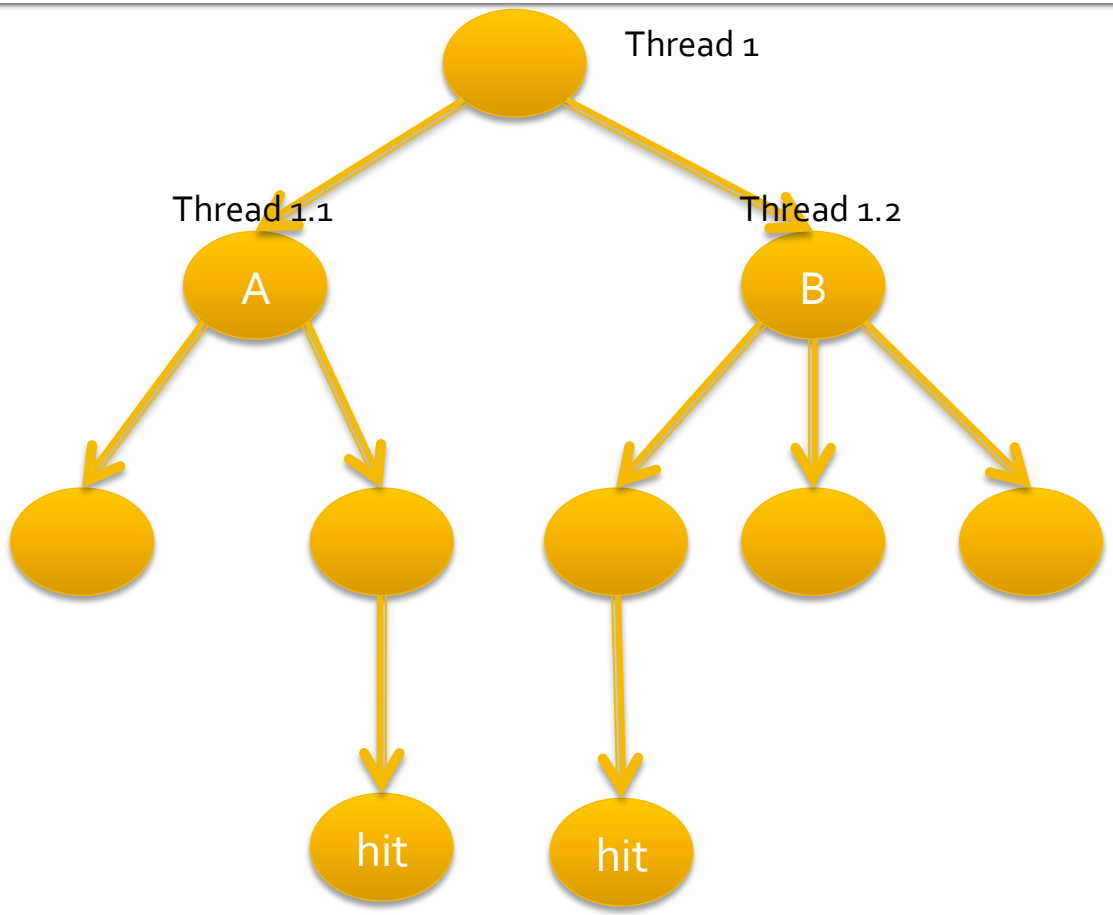
Branch divergence



UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Multi-kernel design



CPU thread queue:

A B

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Mapping accuracy

- Artificial dataset from *C. Elegans* genome
  - 1M 70bp simulated reads with 2% base error rate

Mapping accuracy

	BWA 0.5.8	BarraCUDA
Mapping	89.95%	91.42%
Error	0.06%	0.05%

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

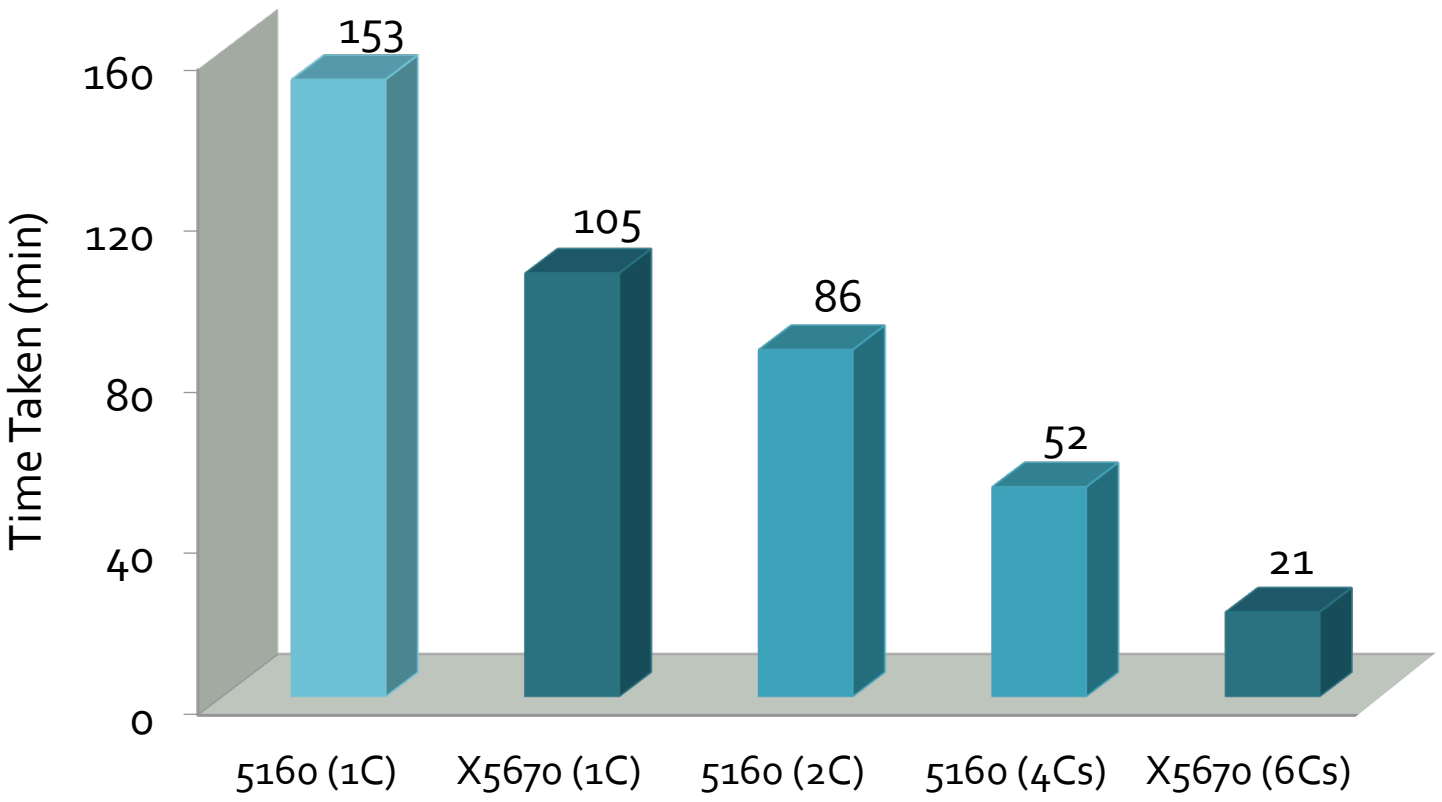
Mapping speed

- Query library: A human 76bp whole genome shotgun library containing 14M reads from the 1000 Genomes Project
- Reference: the Human genome

Hardware configurations

- CPUs:
  - 2x Intel Xeon 5160 (2Cs) @ 3GHz with 8GB DDR2 RAM and fast RAID storage
  - 1x Intel Xeon X5670 (6Cs) @ 2.93GHz with 8GB DDR3 RAM and fast RAID storage
- GPUs:
  - NVIDIA Tesla M2090 /w 6GB GDDR5 RAM

Mapping speed



UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Multiple GPUs

- A shell script to further divide read library into smaller chunks
- Distributes chunks to multiple GPUs

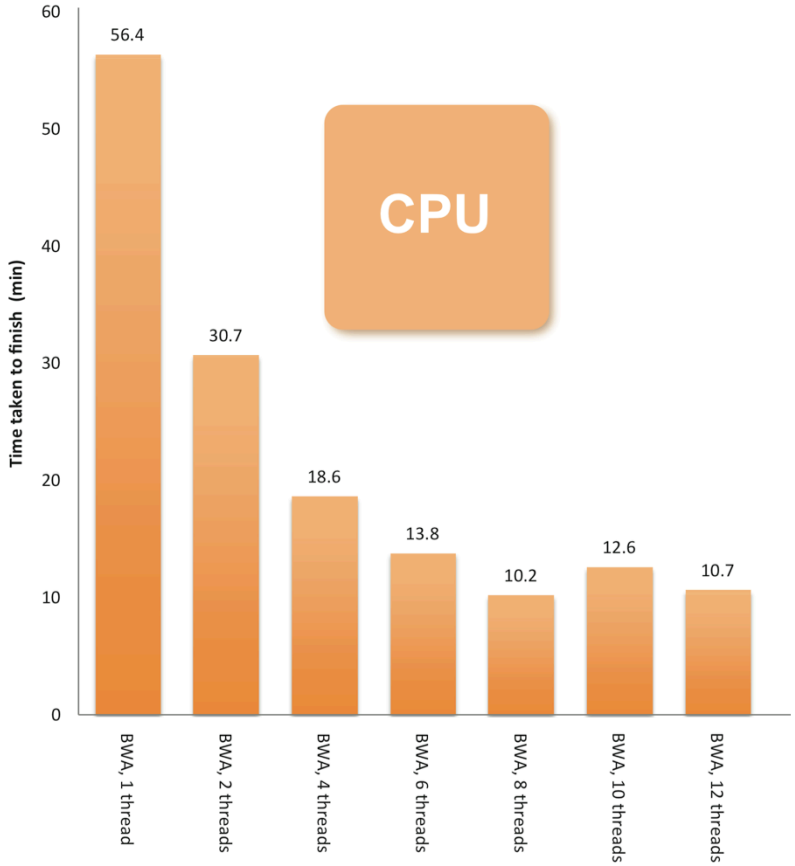
Hardware configurations

- CPUs:
  - 2x Intel Xeon X5670 (6Cs) @ 2.93GHz with 8GB DDR3 RAM and fast RAID storage
- GPUs:
  - 8x NVIDIA Tesla M2050 /w 3GB GDDR5 RAM

Scalability

- Query library: A fly 95bp whole genome shotgun library containing 13.5M reads from the ENA
- Reference: the fly genome

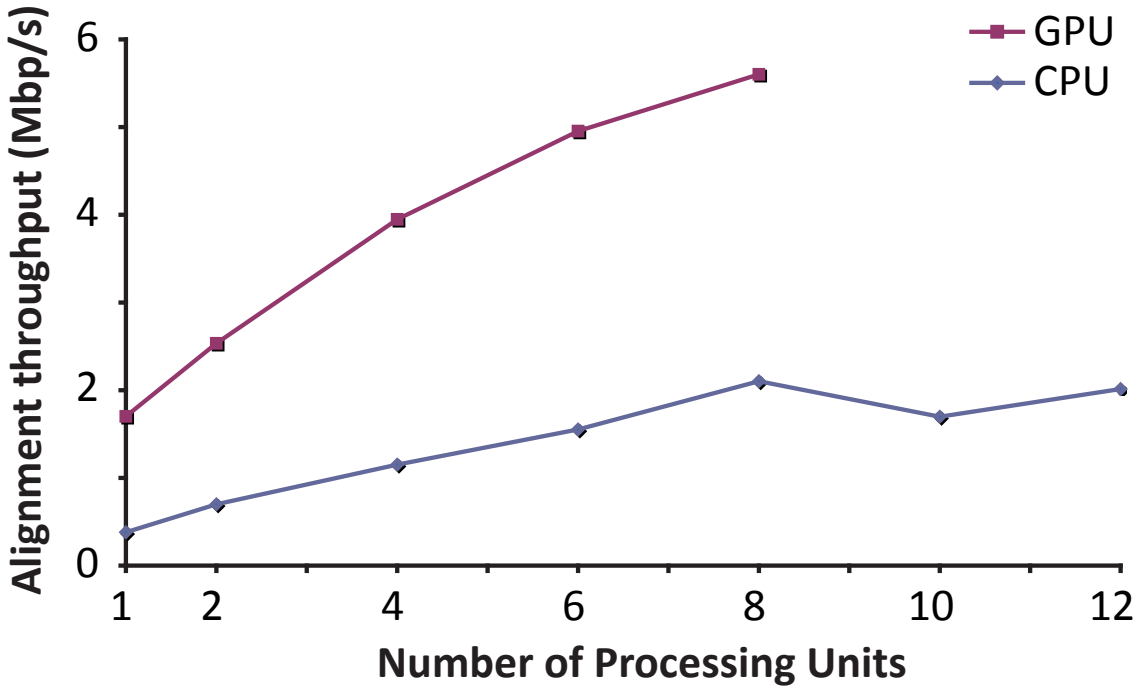
Scalability



UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Scalability



UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Conclusions

- We developed BarraCUDA to map sequence reads onto a genome using GPUs
- The performance from 1 GPU is roughly the same as 6 Xeon CPU cores
- Superior scalability compared to CPUs

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

Future Outlook

- The game is changing
  - CPUs are getting more cores : AMD 16-core Opteron 6200 series
- Many-core platforms are evolving
  - Intel MIC
  - NVIDIA Kepler platform
  - AMD Radeon 7900 series
- OpenCL

Acknowledgements

IMS-MRL, Cambridge
Giles Yeo
Petr Klus
Simon Lam

NIHR-CBRC, Cambridge
Ian McFarlane
Cassie Ragnauth

Whittle Lab, Cambridge
Graham Pullan
Tobias Brandvik

Microbiology, University College Cork
Dag Lyberg

Gurdon Institute, Cambridge
Nicole Cheung

HPCS, Cambridge
Stuart Rankin

NVIDIA Corporation
Thomas Bradley
Timothy Lanfear

MRC Centre for Obesity and Related Metabolic Diseases

EurOCHIP

SEVENTH FRAMEWORK PROGRAMME

EUROPEAN COMMISSION

NHS National Institute for Health Research

NVIDIA

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science

BarraCUDA is now available @:

http://seqbarracuda.sf.net

UNIVERSITY OF CAMBRIDGE
Metabolic Research Laboratories

IMS Institute of Metabolic Science