

A high-resolution photograph of an NVIDIA Tesla GPU card, viewed from a three-quarter perspective. The card is black with a prominent circular cooling fan on the top surface. The NVIDIA logo and the word 'TESLA' are printed on the top. The front edge features a multi-pin connector and a DVI port. The PCIe gold fingers are visible at the bottom.

# BARRACUDA

A High-throughput Sequencing Alignment Software using Graphics Processing Units

Brian Lam

3rd UK GPU Conference, 14<sup>th</sup> Dec 2011

# 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*

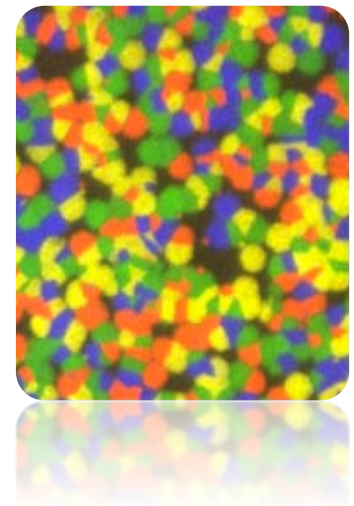
# High-throughput DNA sequencing

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

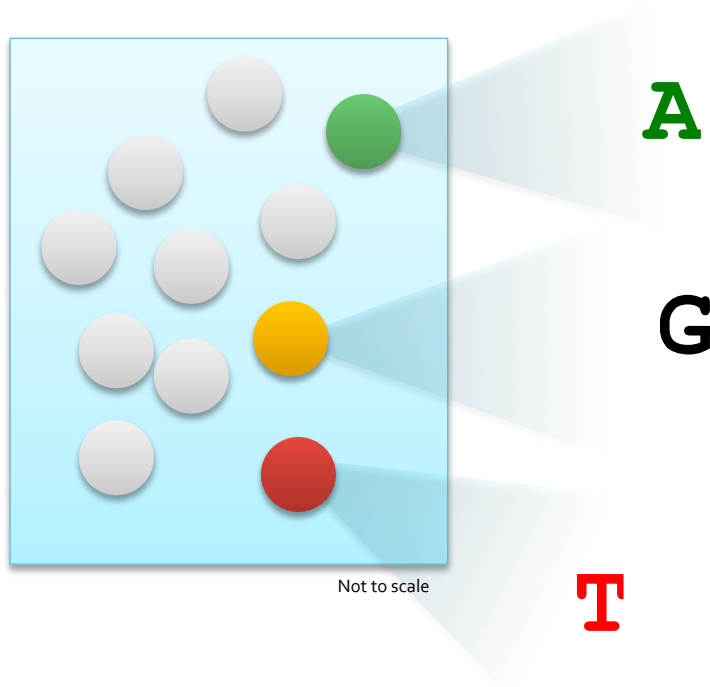
# How it works

- 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 / mm<sup>2</sup>



# Capturing the sequencing signals

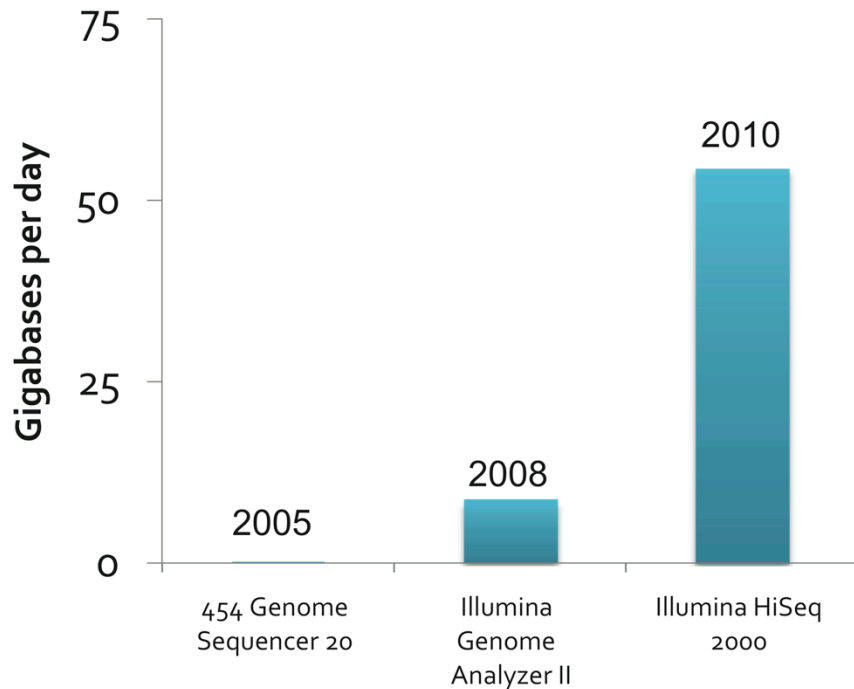


# What do we get from the sequencers

```
barracuda_test_data — less — 162x41
less
@SRR032215.1519 SL-XAT_993091210:3:1:20:610/1
TAGGATTTTTTTTTTCTATTCTGTGAAGAATGTTATTGGGATTTGGAATCTATAGTTTCAGTGCAATCCCTATTA
+
HEHHEHHHHHHGGGCEAHGG@FHGHDGGG-@FDDEB?EH<?GDFDE/F?B=D1C5CB;2<433?-CB=;F=CG#
@SRR032215.1520 SL-XAT_993091210:3:1:20:856/1
ACCTAGCTAATTAGTGGGACTGTCCCGGCACACAGCTAGTCTGCAAAAGACCCGGGAAAGATGGCTTTTTCAAATGCC
+
HHGHHHGHHGHHHHHEHHDHHHHHHCHHEHHHHHHHHHHHHFHHHHHHFHHGEGGHDHGHHGICH=HEHHEFFGGHGH
@SRR032215.1521 SL-XAT_993091210:3:1:20:1688/1
TGATCATCTGTGAGGGGCTCCGGCACAGGGCGATGTAGTCAGGTGAAATGTAGCGAGTATGATCTGACTTTTTCA
+
E?9A5FA:EFFFF==<:CC3;A3,;,<30?E3>C>;9C+;;+79A1414.C);;C2+3?+C844.3?5?#####
@SRR032215.1522 SL-XAT_993091210:3:1:20:1615/1
ACTGTGAGAAAACCTCTTTGTTATGTGAGCATTCAACTCACAGAGTTGAACCTATCTTTTGATTGAGCAGTTTTGAA
+
HHHHHHHHHGHGFHHGHHHHHHHHHHHHHHHHH@GHGHEHEHHHHHGGHHDV.FFF=HHHGHI IHHG9AHDGEGFEB
@SRR032215.1523 SL-XAT_993091210:3:1:20:145/1
TTCTCCTCATTCTCTCCCTTGTCTTCCCGCCACCCCTTCCACTCTCTGGTCTCTTTCATCACTTAAACACTTACCC
+
#####
@SRR032215.1524 SL-XAT_993091210:3:1:20:235/1
CAATCCCAAGAAAATACAAACTGCTATGAGAGAACTATAAACACCTCTATGAAATACGCTAGAAAATCTAGAA
+
HFHH=H0CFHGHGHEHGECHH@HFHFFH=FFFFFHHHHGHDDHDFHHHFGGHHHGEHHH1HBFH9HHHHHH
@SRR032215.1525 SL-XAT_993091210:3:1:20:1700/1
CTCGCTCTGTATCTCAAATATTAGTGTCTCTTCCCTCTGTGTCCAAAATGTACTCATACCCATCTCAGGGAATTC
+
HBHGHHHHHGHHHEHGHGHHGHHGHE=EAFHHH<HBHH3=F=BDHHHGG?ADF7FBFF?CB.D5FFFB@GF=DD
@SRR032215.1526 SL-XAT_993091210:3:1:20:1399/1
GGTTTGTGACTGGGAGCTAGCTGTGGATGGAGGGGAGCAGACAGAGAAAAGAGGATTTATTTAAGCCAACAT
+
HFHHHHGGHDHHHHHHGHHHDHHEHHHHAGHBEHERE;H@HGEC?0<F<FAC?EGDHEHEG=6AB>DF;+?B>A>
@SRR032215.1527 SL-XAT_993091210:3:1:20:224/1
TGTGCTGTCTCAGATCTCACTAGTAAGTGGACTGTGGCCATGCATTCAATTAGCTCTGTTTACAGGGAGATG
+
FHEFHHEHHHFFHDHEHHHEFHDFFBFHCDEFCA+BF@FGFDCCH;B@FEHFEF0E=ACBCAF=FEHFF?H<FH
@SRR032215.1528 SL-XAT_993091210:3:1:21:474/1
TTGCATTCATTCCATTACATTGGGATTGATTCTATTCAACACCCCTTACTCTCCAATTACATTCCATTCCGCGG
+
HHH1HHGHEH1HHDHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGCHGHEHGHHHH8IHG
:Q
```

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

# The throughput of HTS has increased dramatically



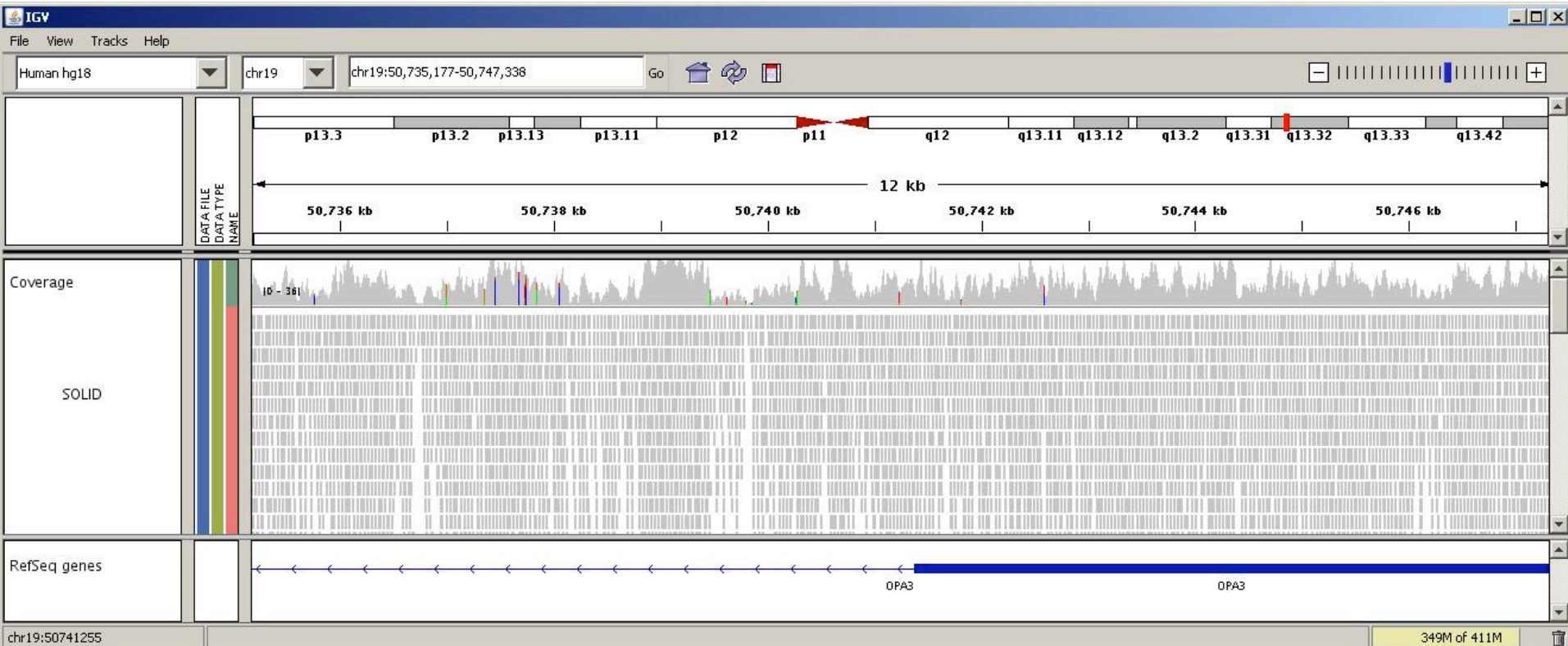
©2010, Illumina Inc. All rights reserved.



# Current bioinformatics pipeline



# Sequence alignment



# 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

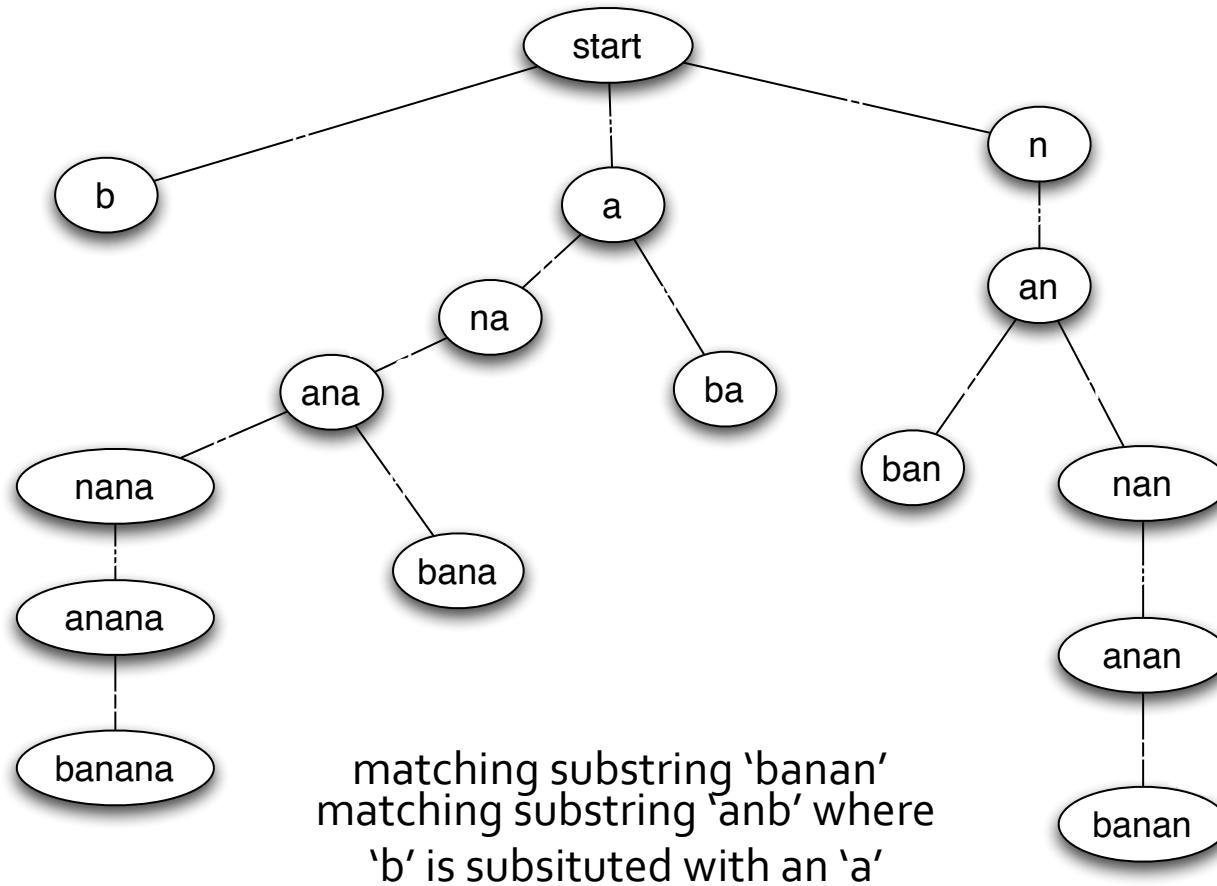
# The BarraCUDA Project

- 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

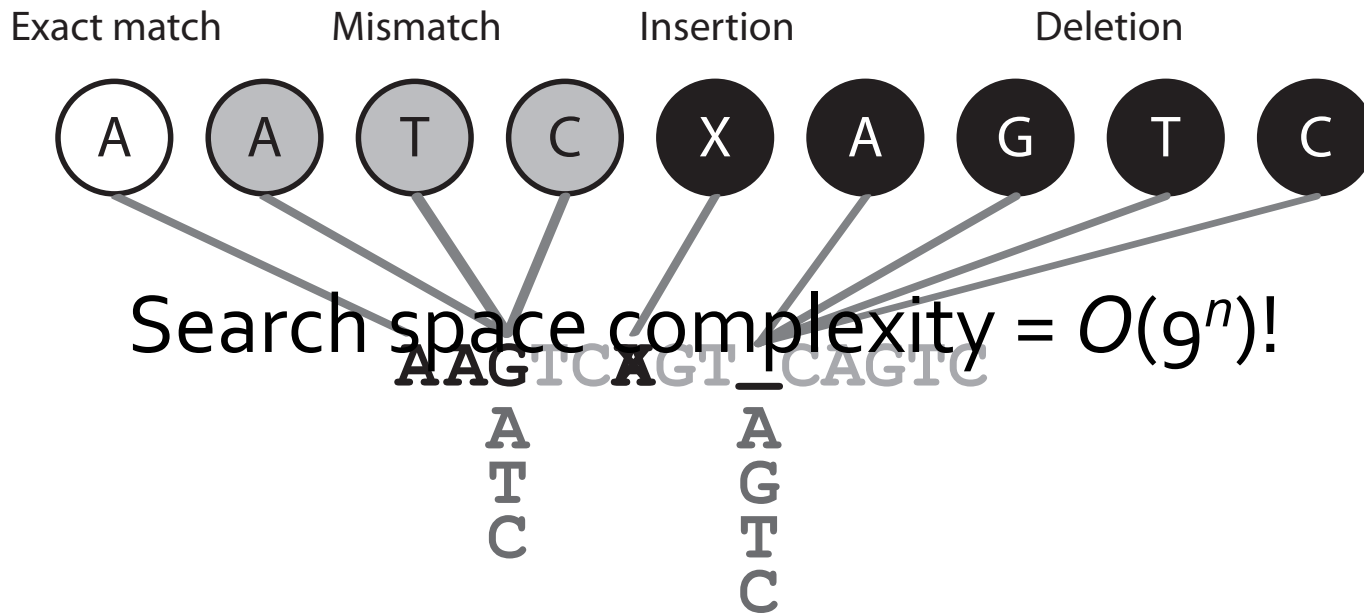
- Originally intended for data compression, performs reversible transformation of a string
- In 2000, Ferragina and Manzini introduced BWT-based index data structure (FM-index) for fast substring matching at  $O(n)$
- Sub-string matching is performed in a tree traversal-like manner
- Used in major sequencing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

# How it works – a backward search algorithm





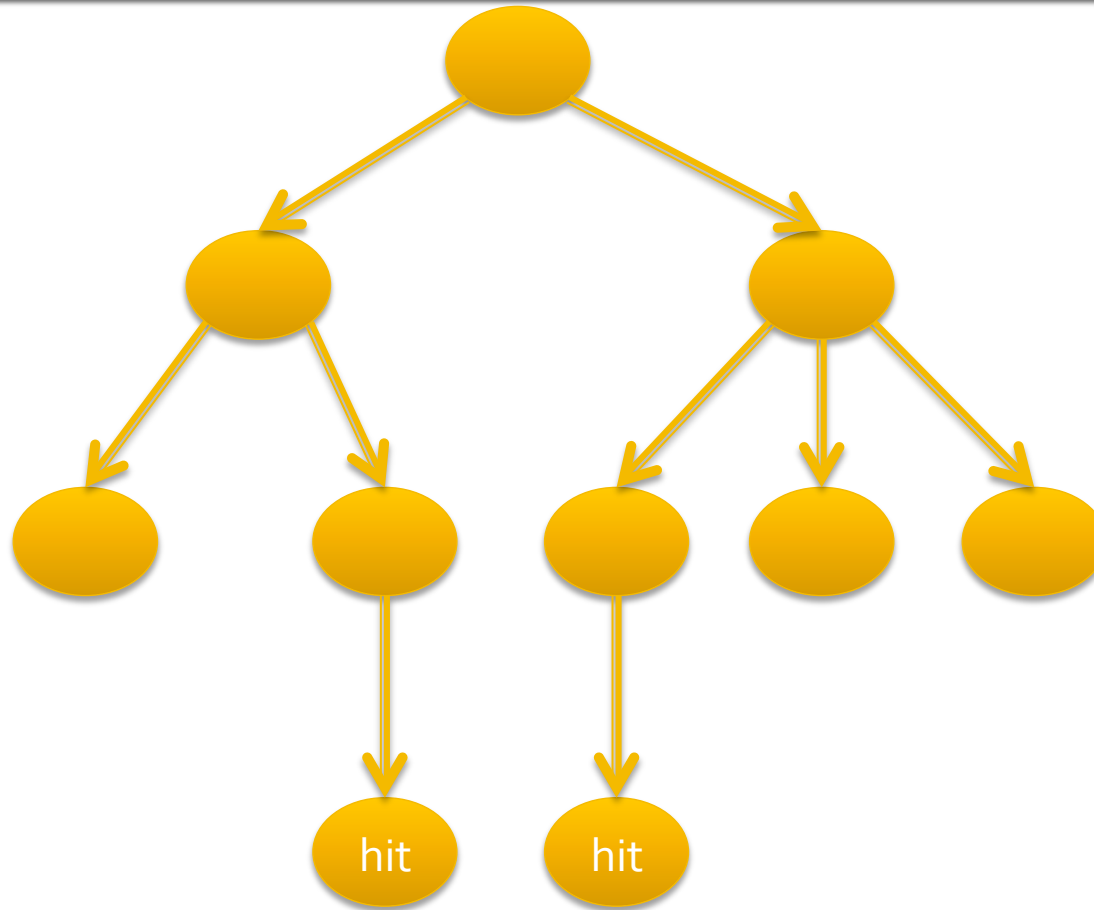
# Inexact matching requires base substitution within the query substring



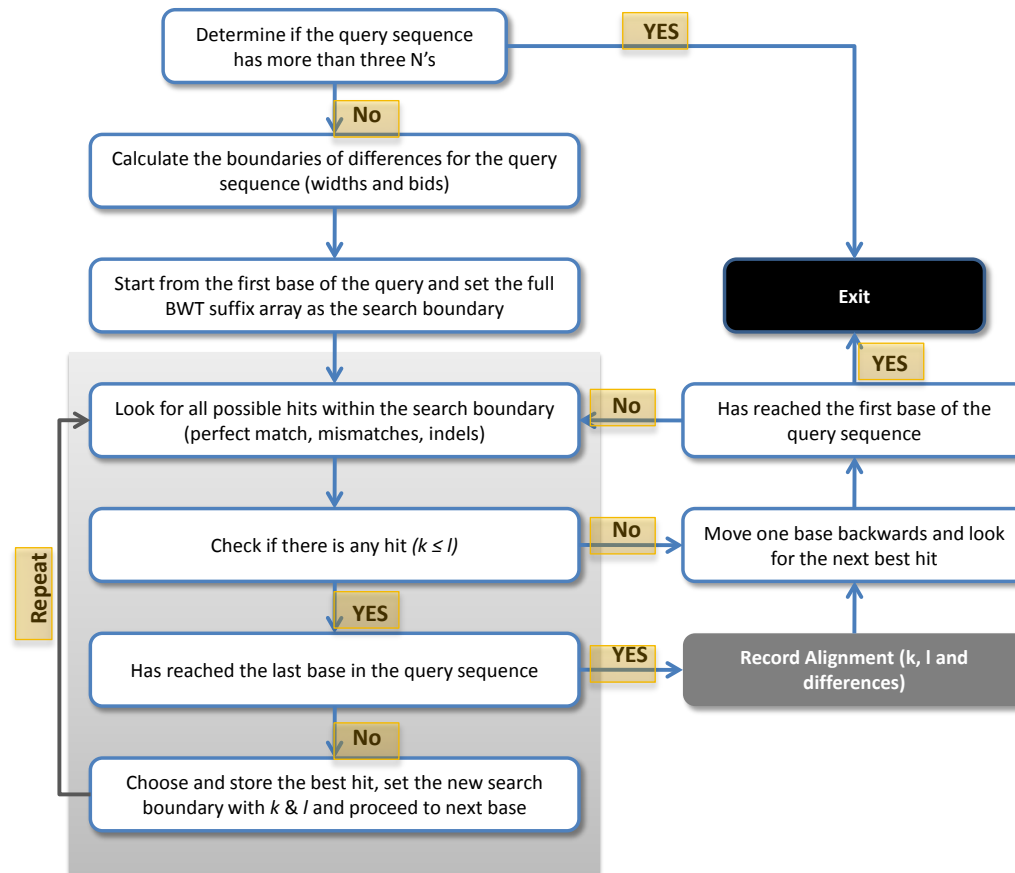
# BarraCUDA started its life as a GPU version of BWA

- 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

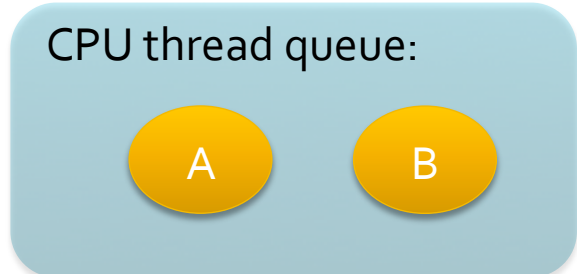
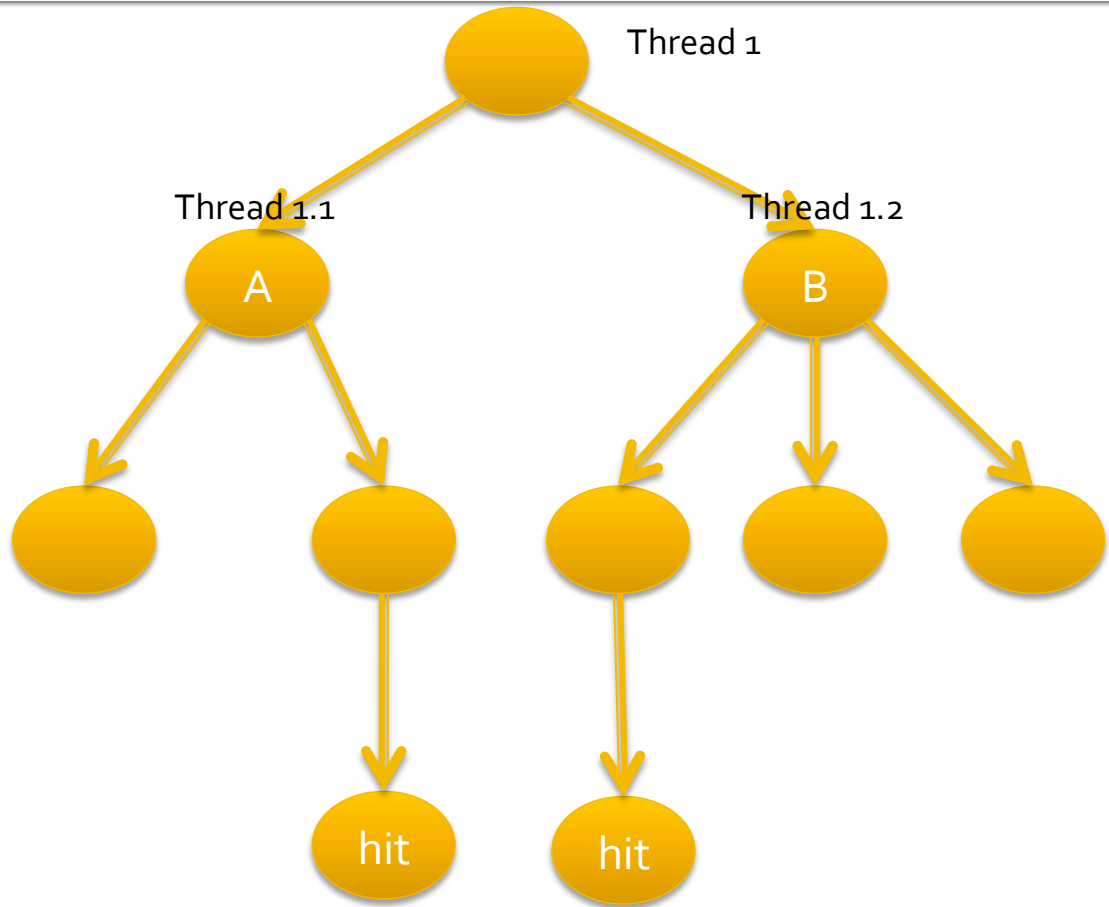
# BarraCUDA uses a depth-first search approach



# Branch divergence



# Multi-kernel design



# 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%

# Mapping speed

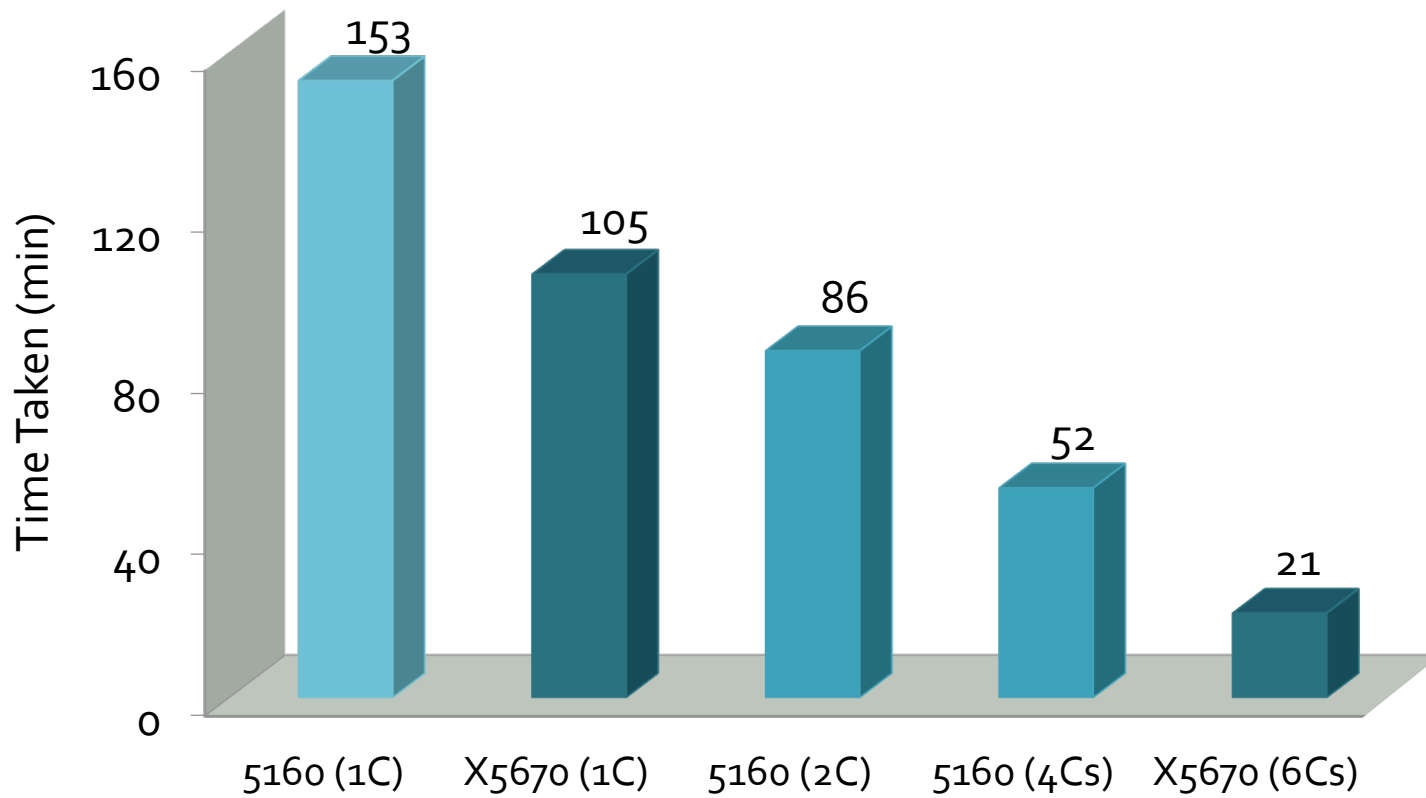
- Query library: A human 76bp whole genome shotgun library containing 14M reads from the 1000 Genomes Project
- Reference: the Human genome



# 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



# Multiple GPUs

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

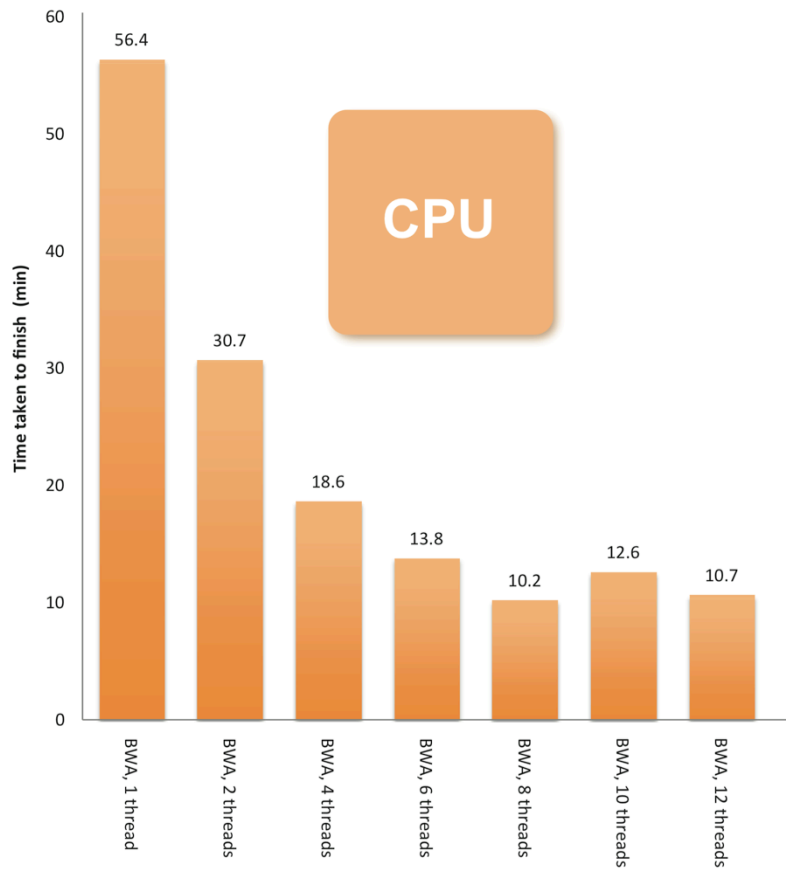
# Hardware configurations

- 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

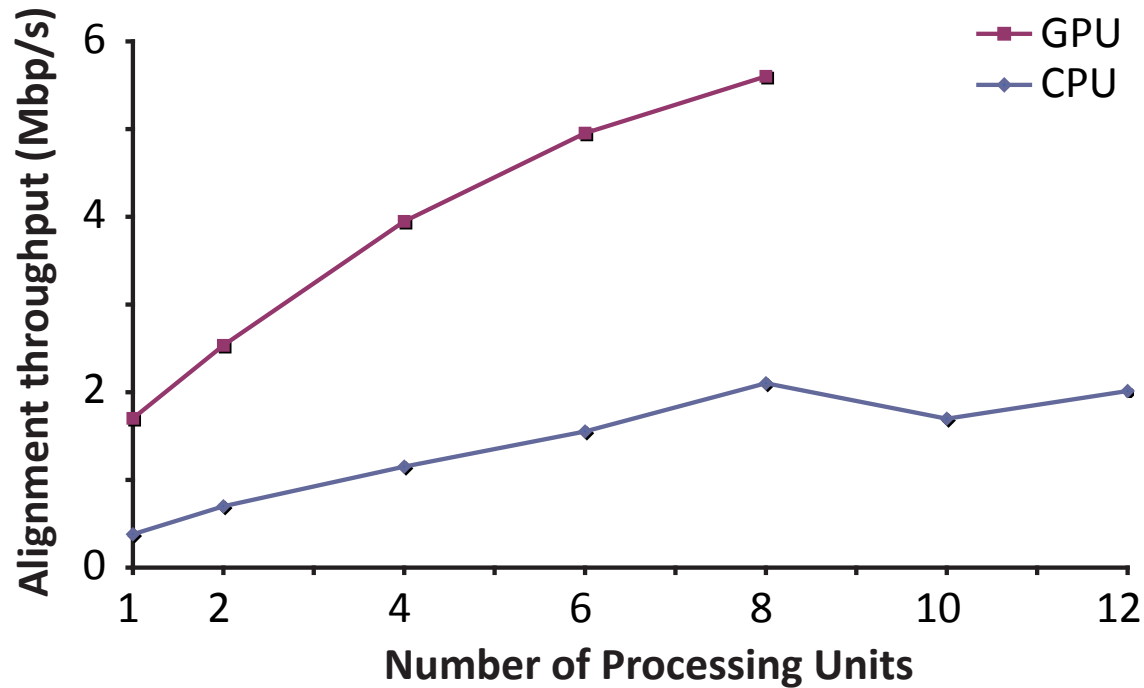
# Scalability

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

# Scalability



# Scalability



# 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



# 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



*National Institute for  
Health Research*



**NVIDIA®**



Institute of Metabolic Science



BarraCUDA is now available @:

<http://seqbarracuda.sf.net>