CUSHAW Software Package: Harnessing CUDA-enabled GPUs for Next Generation Sequencing Read Alignment

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

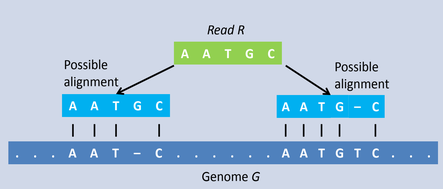
Yongchao Liu and Bertil Schmidt

Institut für Informatik

NGS read alignment

Given a read *R* and a reference genome *G*, NGS read alignment aims to determine *R*'s likely point of origin with respect to *G*.

- Usually exert distance constraint (e.g. edit distance or Hamming distance)
- Usually assume that the true alignment has the optimal alignment score



Seed-and-extend heuristic is the most popular model for NGS read gapped alignment.

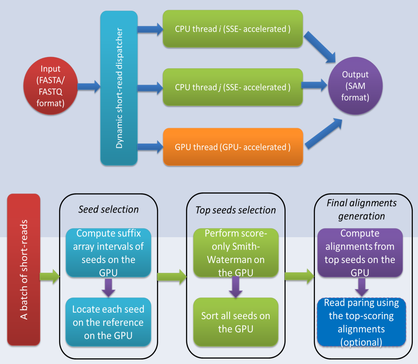
- Seed generation
- Extend and refine seeds
- Produce local/global/semi-global alignments around seeds

Type of Seed		Features	Example Aligners
Fixed-length seeds	Exact *k*-mers	Neither substitutions nor gaps are not allowed	GASSST
	Inexact *k*-mers	Merely allow substitutions	**CUSHAW**, BWA, Bowtie
		Allow both gaps and substitutions	Bowtie2
	Spaced seeds	Only substitutions at predefined positions	BFAST
	q-gram filters	Both Substitutions and gaps are allowed	SARUMAN
Variable-length seeds	Adaptive seeds	constrained by the number of seed occurrences.	LAST
	Long gapped seeds	Limited by a optimal local alignment score threshold	BWA-SW
	Maximal exact match (MEM)	Exact matches that cannot be extended in either direction without allowing a mismatch	**CUSHAW2**, BWA-MEM
Hybrid seeds		Combining multiple types of seeds, e.g. MEM seeds, exact *k*-mer seeds and variable-length seeds	**CUSHAW3**

Gapped alignment on GPUs

Inter-task hybrid CPU-GPU parallelism model

- each CPU thread employs streaming SIMD extension (SSE) instructions
- an instance (i.e. a single process) supports only one GPU, and multi-GPU support can be realized by running multiple instances



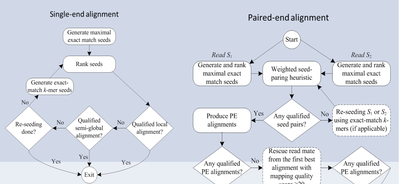
Program workflow of CUSHAW2-GPU

- Seed generation
- Top seeds selection
- Final alignments generation

SOLiD read alignment

Hybrid seeding in CUSHAW3

- maximal exact match seeds
- exact *k*-mer seeds
- variable-length seeds derived from optimal local alignments



Weighed seed pairing heuristic

- q_i is the optimal local alignment score between read S_i (1≤*i*≤2) and the mapping region corresponding to a seed
- *m* is the positive score for an alignment match
- A max-heap data structure is employed with *w* as the key

$$w = \frac{2w_1w_2}{w_1 + w_2}$$

$$w_i = \frac{q_i}{|S_i| \times m}, \quad i = 1,2$$

Abstract

Aligning next generation sequencing (NGS) reads to a genome is often essential to many applications of NGS and has motivated the development of several gapped read aligners as reads become longer. Although usually fast, existing gapped aligners still have difficulties in substantially meeting the requirements of ever-increasing sequence volume on a single multi-core CPU. In this regard, it becomes attractive to resort to accelerators like many-core GPUs.

We present CUSHAW2-GPU to accelerate the CUSHAW2 algorithm using CUDA-enabled GPUs. By aligning both simulated and real reads to the human genome, our aligner yields comparable or better performance compared to BWA-SW, Bowtie2 and GEM. Furthermore, CUSHAW2-GPU with a Tesla K20c GPU achieves significantly speedups over the multi-threaded CUSHAW2, BWA-SW, Bowtie2 and GEM on the 12 cores of a high-end CPU for both single-end and paired-end alignments. In addition, we have presented some features of CUSHAW3, an extension of CUSHAW2 to further improve the alignment quality of base-space reads and offer new support for color-space reads. For color-space alignment, CUSHAW3 is consistently one of the best aligners compared to SHRiMP2 and BFAST.

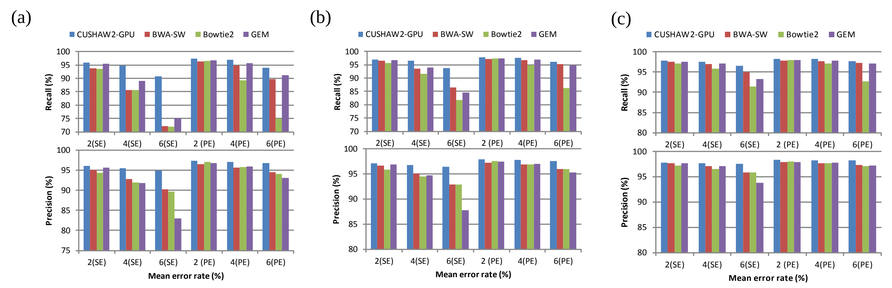
References

1. Yongchao Liu, Bertil Schmidt and Douglas L. Maskell (2012) **CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform**. *Bioinformtics*, 28(14): 1830-1837
2. Yongchao Liu and Bertil Schmidt (2012) **Long read alignment based on maximal exact match seeds**. *Bioinformatics*, 28(18): i318-i324
3. Yongchao Liu and Bertil Schmidt (2013**) CUSHAW2-GPU: empowering faster gapped short-read alignment using GPU computing**. *IEEE Design & Test of Computers*, in press, doi:10.1109/MDAT.2013.2284198
4. Yongchao Liu, Bernt Popp and Bertil Schmidt (2014) **CUSHAW3: sensitive and accurate base-space and color-space short-read alignment with hybrid seeding**. *PLOS ONE*, in press, doi:10.1371/journal.pone.0086869

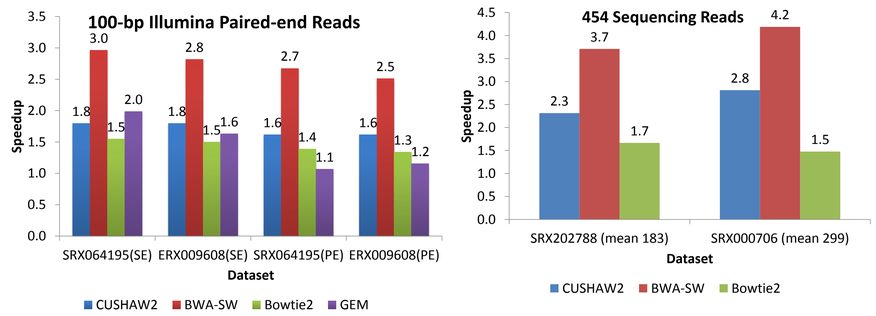
Contact

Yongchao Liu: liuy@uni-mainz.de
Bertil Schmidt: bertil.schmidt@uni-mainz.de
Institute of Computer Science,
University of Mainz
Germany

Performance Evaluation



CUSHAW2-GPU alignment quality on Illumina-like reads: (a) 100 bp; (b) 150 bp; and (c) 250 bp



- A hex-core 12-thread Intel Xeon E5-2620 2.0 GHz CPU and a Tesla K20c GPU
- CUSHAW2, BWA-SW, Bowtie2 and GEM run 12 threads on the CPU
- CUSHAW2-GPU runs on both types of processing units, i.e. CPU and GPU

Runtime breakdown and workload distribution

Assessment	ERX009608		SRX064195	
	SE	PE	SE	PE
Runtime breakdown (in %)				
Read loading	4.6	3.6	4.9	3.6
Seed generation	37.5	29.0	25.5	18.5
Top seed selection	38.4	31.1	49.3	36.6
SW alignment backtracking	1.6	2.7	1.7	3.3
Others*	17.9	33.6	18.6	38.0
Workload distribution (in %)				
#reads processed by the CPU threads	42.0	50.2	40.0	49.2
#reads processed by the GPU thread	58.0	49.8	60.0	50.8

Alignment quality and runtimes on simulated color-space reads

Dataset	Measure	CUSHAW3	SHRiMP2	BFAST
50-bp	Sensitivity	92.13	91.55	88.94
	Recall*	86.28	88.58	81.01
	Recall**	84.72	84.22	81.01
	Time(min)	41	227	160
75-bp	Sensitivity	92.27	92.33	93.44
	Recall*	91.16	91.24	86.14
	Recall**	89.31	88.15	86.14
	Time(min)	20	263	389

*means the recall is calculated from all reported alignments per read and ** means the recall is calculated form the first alignment occurrence per read.

- CUSHAW3 outperforms both SHRiMP2 and BFAST for the 50-bp dataset, while BFAST is the best for the 75-bp dataset, in terms of sensitivity
- CUSHAW3 achieves a speedup of 9.5 (and 11.9) over SHRiMP2 (and BFAST) on average
- For the 75-bp dataset, CUSHAW3 runs 13.5× and 19.9× faster than SHRiMP2 and BFAST, respectively