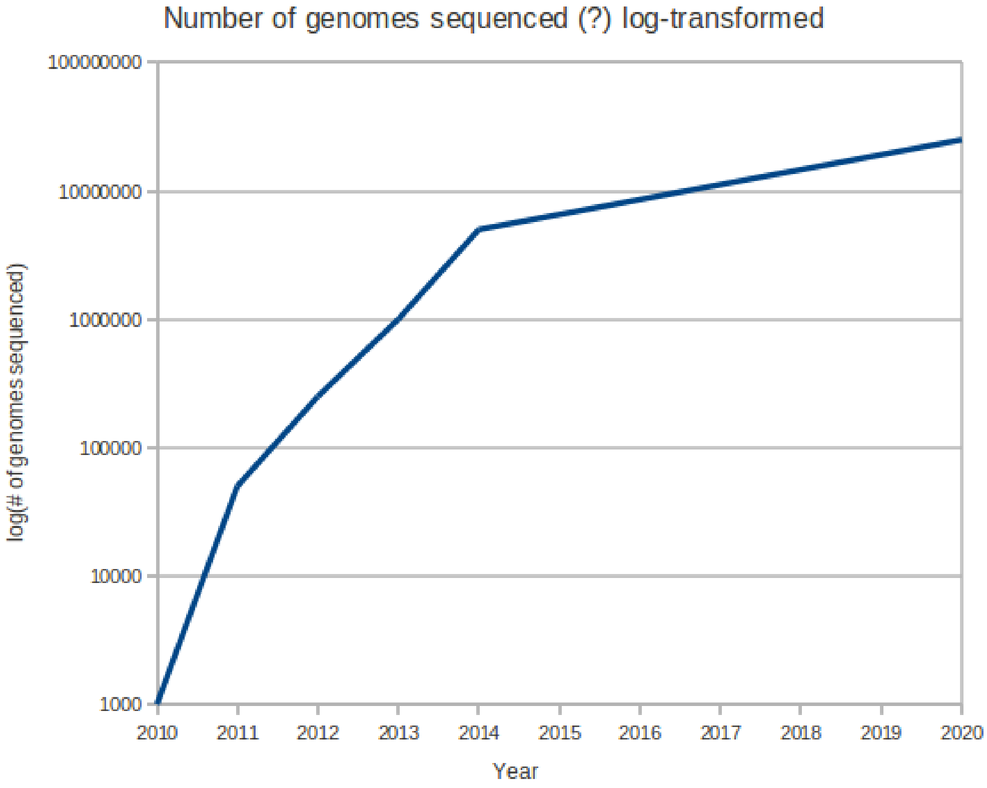
nature|methods
Application Notes

SeqNFind®: A GPU Accelerated Sequence Analysis Toolset Facilitates Bioinformatics

D. Andrew Carr, Christine Paszko, and Donald Kolva

The SeqNFind® genomic sequence analysis toolset accelerates bioinformatics research addressing the critical need for extremely fast and complete sequence analysis tools to handle the huge data volumes generated by Next Generation Sequencing (NGS) platforms (i.e., Roche 454, Illumina® and the Ion Torrent®. Genomic researchers are currently faced with massive amounts of genomic data which are rapidly being generated and require analysis, "In the past year there has been a boom in molecular genetics technology (figure 1) and this has lead to an unprecedented amount of genomics data"[1, 2]. SeqNFind® is the first bioinformatics system of its kind that leverages the power and performance of Graphical Processing Units (GPUs) with optimized sequence alignment algorithms to increase throughput.

Figure 1.



Data supplied by Genome Quest Company[6].

The ability to accurately and quickly analyze, manage and store this data is becoming more important than ever. The massive challenge facing the NGS community is accurately processing data with a manner fast enough to keep up with the current production of data. To date computational processing throughput has been the major hurdle, more complete algorithms such as Smith-Waterman have been too slow to use. To circumvent this hurdle, bioinformaticians employ heuristics to reduce the search space. It is well known that heuristics, although much faster, are likely to return incomplete results.

For example, scientists that have been leveraging BLAST[3], Basic Local Alignment Search Tool, an algorithm for comparing primary biological sequence information may be overlooking significant findings in their data, due to the current lack of tools to provide the needed granularity[4]. While BLAST is faster than Smith-Waterman, a well-known algorithm for performing local sequence alignment, it cannot "guarantee the optimal alignments of the query and database sequences" whereas, the optimality of Smith-Waterman" ensured the best performance on accuracy and the most precise results"[4] at the expense of time and computer power.

Unlike most of the traditional sequence analysis tools, the SeqNFind® system leverages commodity hardware, the massively parallel architecture powered by GPUs to carry out 15,000 to millions of small sequence comparisons simultaneously, and provides a fully scalable solution offering scientists a turn-key solution for sequence analysis. GPUs are causing a paradigm shift to faster and more precise results; they offer a 112-fold increase in the number of processors over traditional CPU (Computer Processing Unit). In addition to extreme speed-ups, GPUs also offer significant space and energy savings (figure 2). Prior to GPUS, Field-programmable Gate Arrays (FPGAs) could be utilized however; some of the challenges with these include much higher costs (as they are not commodity hardware), difficult development, and they are not “future-proof”. The SeqNFind® system, based on NVIDIA CUDA™, is a user friendly (figure 3), open platform that allows scientists to develop additional applications, extend the toolset and leverage other software applications. The exploitation of the GPU hardware with optimized algorithms has allowed researchers to begin exploring genomic data in a novel and affordable manner.

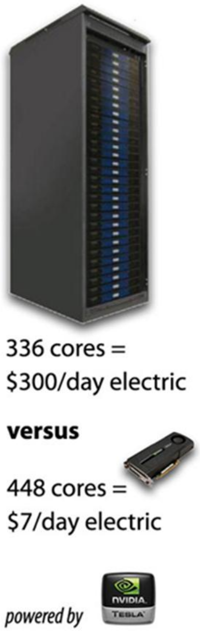
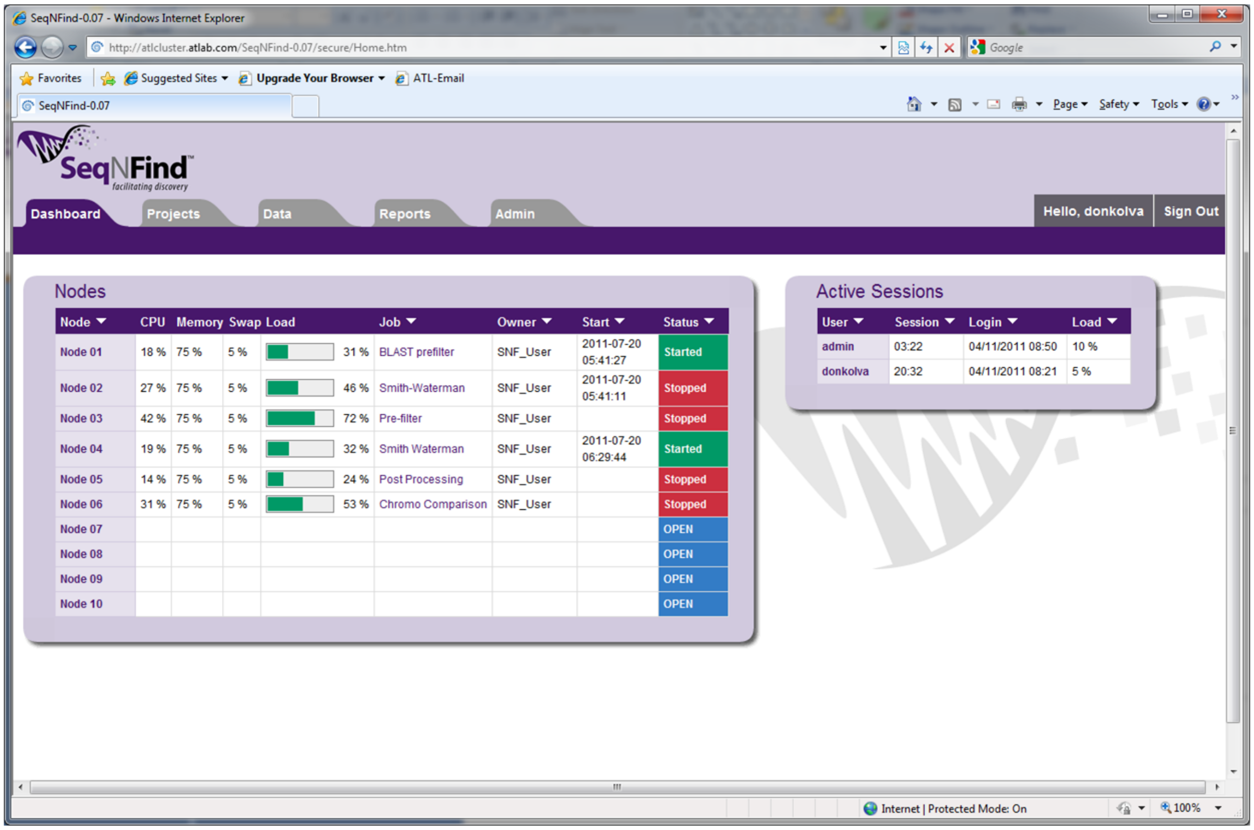


Figure 2. A comparison of the number of cores of a traditional CPU system to a GPU cores and the associated size and energy savings

Current Applications

The SeqNFind® Smith Waterman tool allows examination of local alignments at every location within a genome. This completeness has lead to new methods of analysis, that have already lead to a very significant discovery of some of the challenges of using chip arrays and exploring in greater detail probe target cross-hybridization, understanding hairpin loops and providing an explanation for false positives[5]. Prior to the use of SeqNFind® tools researchers were limited in the scope of the data they could explore in a reasonable time. Earlier attempts at this study had required months of compute time on accelerated hardware. SeqNFind® allowed genomic results to be obtained within days (rather than months or years on traditional CPU systems), thereby allowing rapid iteration, validation and the development of new tools and ideas. Other ongoing projects are currently leveraging SeqNFind® to explore short read sequence data, look at phylogenetics, and to validate and complete the data holes left by heuristic algorithms.

Figure 3. Image of the SeqNFind® user dashboard.



Nodes

Node	CPU	Memory	Swap	Load	Job	Owner	Start	Status
Node 01	18 %	75 %	5 %	31 %	BLAST prefilter	SNF_User	2011-07-20 05:41:27	Started
Node 02	27 %	75 %	5 %	46 %	Smith-Waterman	SNF_User	2011-07-20 05:41:11	Stopped
Node 03	42 %	75 %	5 %	72 %	Pre-filter	SNF_User		Stopped
Node 04	19 %	75 %	5 %	32 %	Smith Waterman	SNF_User	2011-07-20 06:29:44	Started
Node 05	14 %	75 %	5 %	24 %	Post Processing	SNF_User		Stopped
Node 06	31 %	75 %	5 %	53 %	Chromo Comparison	SNF_User		Stopped
Node 07								OPEN
Node 08								OPEN
Node 09								OPEN
Node 10								OPEN

Active Sessions

User	Session	Login	Load
admin	03:22	04/11/2011 08:50	10 %
donkolva	20:32	04/11/2011 08:21	5 %

The SeqNFind® solution offers an affordable workstation for the single researcher and a cluster configuration for the institutions that can leverage massively paralleled commodity hardware adding additional nodes as needed, as SeqNFind® employs the NVIDA TESLA chipset in the hardware boxes- workstation or cluster nodes, and additional TESLA GPUs can be added to keep the system current, thereby future-proofing the technology investment. The algorithms and toolset that are employed include; SeqNFind® Reference Assembly, Smith-Waterman, BLAST, Needleman-Wunsch, HMM, DeNovo Assembly and RNA-Seq.

References:

1. Berger, S. A., D. Krompass, and A. Stamatakis. (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. Syst. Biol. 60 (3):291-302
2. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141

3. Altschul, Gish, Miller, Myers. (1990), Lipman *Basic local alignment search tool*, J. Mol. Biol., Band 215, S. 403–410
4. Wieds, Gustav (2007). "Bioinformatics explained: BLAST versus Smith-Waterman". CLCBio. http://www.clcbio.com/index.php?id=1098.
5. D. A. Carr, S. Koshnevis, D. Kolva, and J. W. Weller (2011) Flanking Sequence Effects on Oligonucleotide Hybridization. ISMB, Vienna, Austria N29
6. p://blogs.discovermagazine.com/gnxp/2010/07/genomic-liftoff/

Authors

Dr. D. Andrew Carr[1], Dr. Christine Paszko[1] & Donald Kolva[1]

1. Accelerated Technology Laboratories, Inc. 496 Holly Grove School Road West End, NC 27376 US