Better Computing for Better Bioinformatics

The World’s First Hybrid-Core Computer.

George Vacek, PhD, MBA
Director, Convey Life Sciences
gvacek@conveycomputer.com
www.conveycomputer.com/lifesciences/
Acknowledgements

• Convey
  – Kirby Collins
  – Wesley Hart
  – Mark Kelly
  – David Soper
  – John Amelio
  – Mike D’Jamoos
  – Mike Ruff
  – Tony Brewer

• BWA
  – Heng Li, The Broad

• Velvet
  – Daniel Zerbino, UCSC

• Virginia Bioinformatics Institute
  – Bob Settlage
  – Lauren McGiver

• Joint Genome Institute
  – Alex Copeland
  – Alex Sczyrba
  – Zhong Wang

• Monsanto Company
  – Yili Chen
  – Xuefeng Zhou

• The Jackson Laboratory
  – Glen Beane
Agenda

• **Better Bioinformatics**
  – High Performance *de novo* Assembly
  – Screening Reads Instead of Contigs
  – High Throughput Resequencing

• **Better Computing**
  – Convey Computers
  – Hybrid-Core Computing
Convey’s Hybrid-Core Server Delivers

- **Higher Performance**
  - 5x to 25x application gains

- **Energy Saving**
  - Up to 90% power reduction

- **Easy to use, program, manage**
  - Standard Linux ecosystem
  - Management / Scheduling
  - Programming environment

“Speed and power consumption were our top reasons for selecting the Convey system.”

Dr. Guilherme Oliveira, Director
Center for Excellence in Bioinformatics
Reduced Memory Usage, Accelerated Performance - Enables Large Genomes

- 5.4x speed up depends on
  - Data set size
  - Kmer space complexity
- RAM reduced 79%
  - Data types / structures
  - Automated roadmap partitioning
- 1.9x Power Performance

HC-2: 2 Intel X5670 2.93GHz processors (12 cores total), stripe 4 @ 600GB SATA disks 96GB DDR3 (host), 16GB SG (coprocessor)
X86: host only

“Convey’s GraphConstructor offers a new approach to help researchers ... to achieve better assemblies or look at bigger jobs such as metagenomic or mammalian genome samples”
Daniel Zerbino, author of Velvet
Convey GraphConstructor for *de novo* Assembly

- Tackle previously impractical genomes
- Higher quality assemblies
- Lower cost

- Interface for Velvet/Oases
- Stability, ease of use, optimized workflow

- Very fast Kmer Counter
  - parameter optimization based on roadmap statistics
  - select best kmer length and coverage cutoff

“Convey is solving a big problem here – *de novo* assembly has been very difficult... Convey has made a significant accomplishment!”

*Dr. John Castle, head of Bioinformatics/Genomics, University of Mainz, TrOn*
SCREENING READS INSTEAD OF CONTIGS
Quickly Identify Reads Associated with Proteins of Interest

- **Translated Search with Smith-Waterman**
- **7.2x faster than BLASTx**
- **2.9x more matches**
  - SWSearch 1081219 hits
  - BLASTx 372344 hits
- BLAST heuristic filter

<table>
<thead>
<tr>
<th>Query</th>
<th>Subject</th>
<th>resemblance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AARTPKPTAPDSPEMMRG</td>
<td>AARLPAPTGPGPFAGRG</td>
<td>153</td>
</tr>
<tr>
<td>PGGTFSSSP</td>
<td>PGGTLFSTXP</td>
<td>1095</td>
</tr>
</tbody>
</table>

1M Illumina Reads Against 5K Patented Proteins (pataa)

HC-2ex: 192GB (host), 64GB (coproc), stripe 4 @ 600GB SATA
Dell r610: 2 Intel X5680 3.33GHz processors (12 cores total), 96GB of 1333MHz DDR3 memory, stripe 3 @ 146GB SAS
HIGH THROUGHPUT RESEQUENCING
Workflow Performance for Human

- **BWA 0.5.10 workflow**
  - 2 @ aln + sampe
  - 8.8x - 9.4x over x86
  - 62 - 67 K Reads/Sec

- **Reference G1k v37**
  - 3.1 G bases

- **Reads**
  - HG00124 SRR189815_(1,2)
  - 242 M reads, ~100 bp
  - 24.7 G bases

---

X86: host only from HC-2ex; Intel X5670 2.93GHz processors (12 cores total), stripe 4 @ 600GB SATA disks
HC-1: 128GB (host), 64GB (coproc), stripe 2 @ 1TB SATA disks
HC-1ex: 128GB (host), 64GB (coproc), stripe 2 @ 1TB SATA
HC-2: 96GB DDR3 (host), 16GB SG (coprocessor)
HC-2ex: 192GB DDR3 (host), 64GB SG (coprocessor)
Further Workflow Optimization
Integrated BAM format generation

- **SAM to compressed BAM**
  - `samtools -bS` vs. `integrated -b`

- **3.9 - 9.8x speed up**

- **Even greater savings for slow file systems**

- **Reference G1k v37**
  - 3.1 G bases

- **Paired-end Reads**
  - SRR002813, SRR002987, SRR002968

---

**HC-2ex:** 192GB DDR3 (host), 64GB SG (coprocessor), stripe 4 @ 600GB SATA disks

9/6/12
The Jackson Laboratory

- Mutations vary in size
  - E.g. translocation breakpoint
  - Want reads that span breakpoint
- Run BWA with varying parameters
  - Get more of these mutations
- Too slow on 32-core servers
- HC-2ex is 11.3x faster
  - Afford to adjust parameters
  - Quickly perform multiple runs
  - Achieve better results

“... GPUs weren’t a good fit for alignment... the performance isn’t that compelling. Other FPGA system vendors didn’t have the number of tools Convey does or the system wasn’t as easy to use. Also a developer community is evolving around the Convey systems where we could share third-party tools.”

Glen Beane
The Jackson Laboratory

BWA 0.5.10
X86: 4 x 8-core AMD Magny Cours 2.4GHz Opteron
HC-2ex: 2 x 6-core Intel X5670 2.93GHz, coprocessor

9/6/12
CONVEY HYBRID-CORE COMPUTING
BWA Personality

- Implemented in hardware on coprocessor FPGAs
- Highly parallel—up to 2,048 simultaneous alignment operations
  - 64 alignment units each operate on 32 sequences simultaneously
- Leverages Convey HC highly parallel memory
## Bioinformatics Applications on HC

<table>
<thead>
<tr>
<th>Organization</th>
<th>Application</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convey Bioinformatics Suite</td>
<td>BWA</td>
<td>Reference Mapping</td>
</tr>
<tr>
<td>Convey Bioinformatics Suite</td>
<td>Velvet/CGC</td>
<td>De Novo Assembly</td>
</tr>
<tr>
<td>Convey Bioinformatics Suite</td>
<td>Kmer Counter</td>
<td>Read Analysis for Assembly</td>
</tr>
<tr>
<td>Convey Bioinformatics Suite</td>
<td>SWSearch</td>
<td>Smith-Waterman Search</td>
</tr>
<tr>
<td>Convey Bioinformatics Suite</td>
<td>BLAST(p,x)</td>
<td>Protein Database Search</td>
</tr>
<tr>
<td>CLC bio</td>
<td>CLC Genomics platform</td>
<td>Analysis, workflows, visualization</td>
</tr>
<tr>
<td>Michigan Technological University</td>
<td>PCAP</td>
<td>Overlap-based Assembly</td>
</tr>
<tr>
<td>University of California San Diego</td>
<td>InsPect</td>
<td>Protein Assembly with PTMs</td>
</tr>
<tr>
<td>BlueSpec</td>
<td>Memocode</td>
<td>Burrows-Wheeler Aligner</td>
</tr>
<tr>
<td>Iowa State University</td>
<td>RMAP, Shepard</td>
<td>Short-read Mapping</td>
</tr>
<tr>
<td>Technical University Crete</td>
<td>BLASTn</td>
<td>Nucleotide Sequence Search</td>
</tr>
<tr>
<td>University of California Los Angeles</td>
<td>Fluid Registration</td>
<td>Medical Imaging</td>
</tr>
<tr>
<td>University of California Riverside</td>
<td>BowTie/FHAST</td>
<td>Burrows-Wheeler Aligner</td>
</tr>
<tr>
<td>University of South Carolina</td>
<td>Mr Bayes</td>
<td>Phylogenetics</td>
</tr>
<tr>
<td>Boston University</td>
<td>BLAST(p,x)</td>
<td>Protein Database Search</td>
</tr>
<tr>
<td>Free University of Berlin</td>
<td>SeqAn</td>
<td>Sequence Analysis Library</td>
</tr>
<tr>
<td>Technical University Darmstadt</td>
<td>GROMACS</td>
<td>Molecular Dynamics</td>
</tr>
<tr>
<td>Bielefeld University</td>
<td>SARUMAN</td>
<td>Short-read Mapping</td>
</tr>
<tr>
<td>University of Paderborn</td>
<td>Suffix Tree</td>
<td>Short-read Mapping</td>
</tr>
<tr>
<td>University of Washington</td>
<td>BFAST</td>
<td>Short-read Mapping</td>
</tr>
<tr>
<td>Virginia Bioinformatics Institute</td>
<td>Various</td>
<td>Mol Dynamics, Bioinformatics</td>
</tr>
</tbody>
</table>
High Throughput Bioinformatics

• **In-house development and collaborations**
  – Customers and partners
  – Software vendors
  – Instrument manufacturers
  – Cloud services

• **Addressing many facets of bioinformatics**
  – primary analysis
  – de novo assembly and reference mapping
  – sequence alignment and search
  – annotation, other downstream analysis

• [www.conveycomputer.com/lifesciences/](http://www.conveycomputer.com/lifesciences/)