

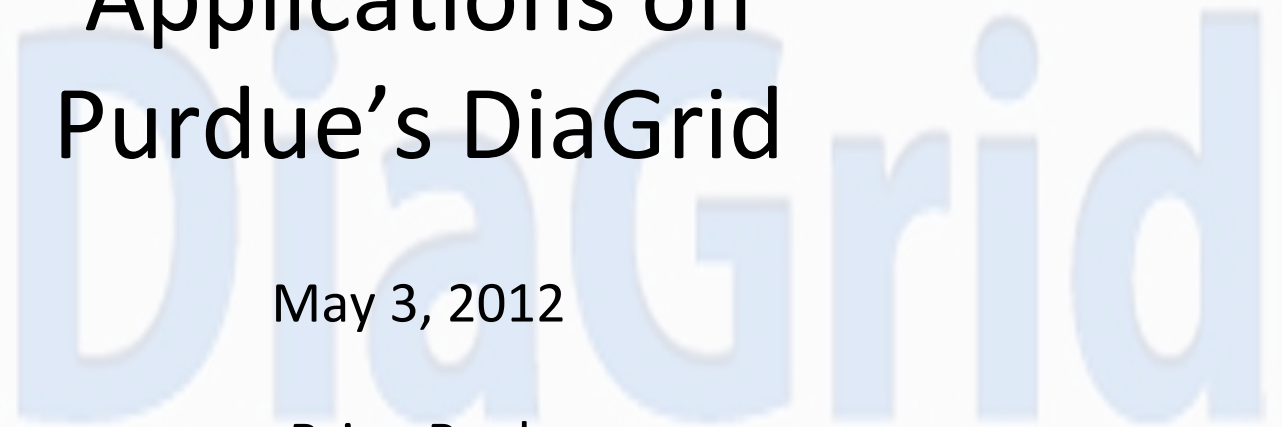
BLAST and Bioinformatics Applications on Purdue's DiaGrid

May 3, 2012

Brian Raub

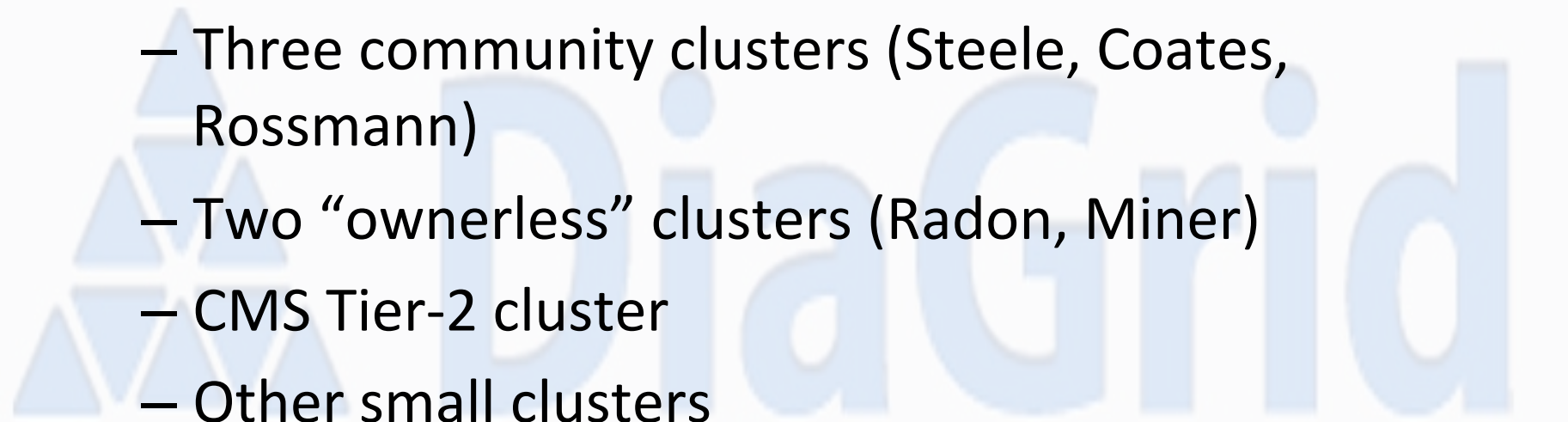
Purdue University

braub@purdue.edu



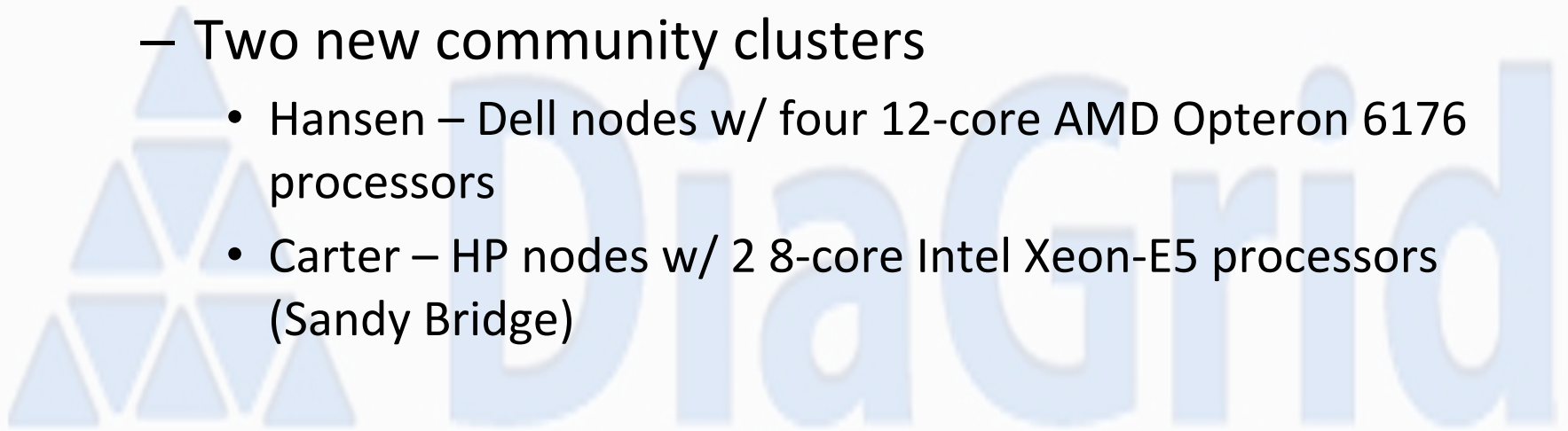
Where were we?

- Over 37 kilocores across campus
 - Three community clusters (Steele, Coates, Rossmann)
 - Two “ownerless” clusters (Radon, Miner)
 - CMS Tier-2 cluster
 - Other small clusters
 - Instructional labs and academic departments



... and what about now?

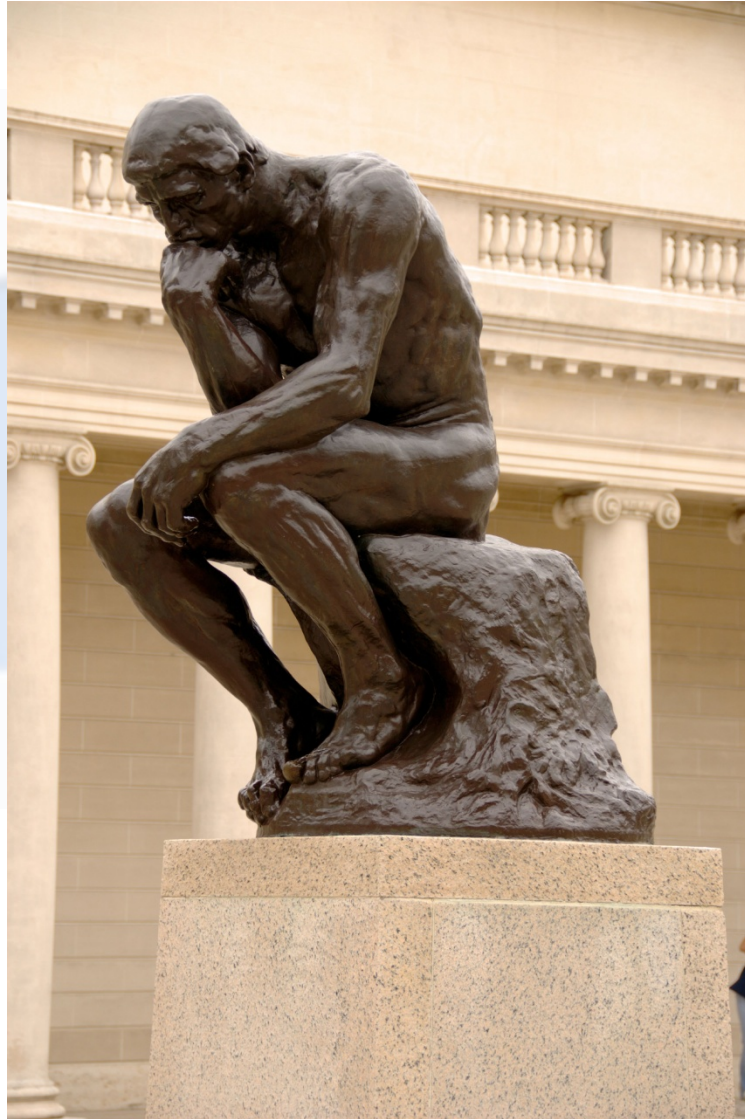
- Nearly 50 kilocores across campus!
 - Two new community clusters
 - Hansen – Dell nodes w/ four 12-core AMD Opteron 6176 processors
 - Carter – HP nodes w/ 2 8-core Intel Xeon-E5 processors (Sandy Bridge)
 - Carter ranks 54th in the latest Top500.org list for fastest supercomputers
 - Carter is the nation's fastest campus supercomputer



DiaGrid?

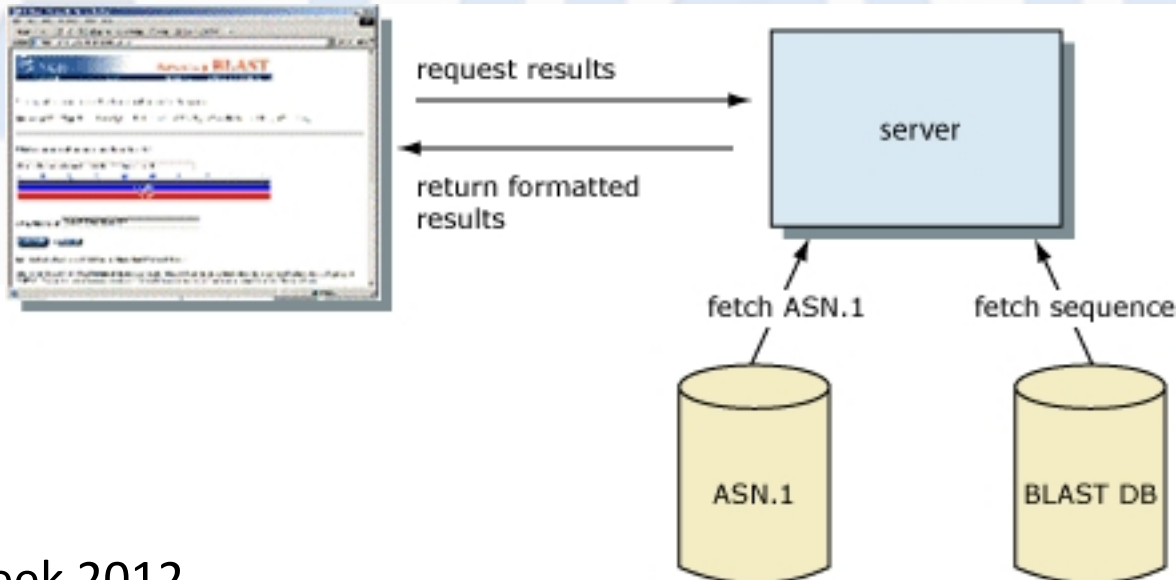
- A large, high-throughput, distributed computing system
- Using Condor to manage jobs and resources
- Purdue leading a partnership of 10 campuses and institutions
 - University of Wisconsin, Notre Dame and Indiana University to name a few
- Including all Purdue (and other campus) clusters, lab computers, department computers, desktop, totaling 60,000+ cores

Ok, cool... Now what?



rid

- Comparing nucleotide or protein sequences
 - String and Substring pattern matching
- National Center for Biotechnology Information (NCBI)



Why remake something?

- Input file size limitations (5MB, 10MB, etc.)
- # of sequences for comparison
- Timeliness
- Ease of use

DiAGrid

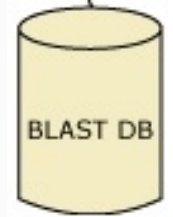
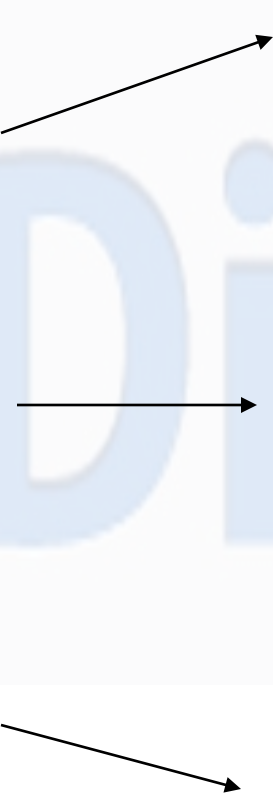
- BLAST is highly parallelizable
 - No one sequence result depends on another (GREAT!!!)
 - Split input file with trusty friend AWK
 - Build a Condor DAG to maintain all jobs
 - Never more than 1500 individual jobs

Input File

```
>comp4_c0_seq1 len=401 ~FPKM=247.6 path=[0]
TTTTTATTGGTAAATAAATATGAGTGGAGTATATAAAGATGGAGAGTATGCGGAG
CGACGGCGAAGGATGAGTAATTTGATGGAGTGGATGGAAGTTTTGTTTTGCGGAAGA
TTTTAATTTTTTAAATATATTTGATTTGGTTTGAATAGATTTGAAATATATG
TTATAGTATATAAAGAGATGATTAAGGATTAATAAGTATATGTTGTAAGTTAG
TAAAGAAATTTTTAGAATTAAGAATATGAGTTTATGTTAAGTCTGGTAAAAGATTGT
TTTAAATGGTTTGGTAAAGATATCGGTTAATGTTTTTTGGTGGTTAATATATGTAATT
TCGATATTTTTTTTTTGTTAGATCGGAAGCTCGTATGCC

>comp7_c0_seq1 len=470 ~FPKM=611.4 path=[11]
TCACCATTTTTTTTCAAGCAGAAATGCGGCATACGAGCTCTTCGCGATCTTTAAACTTAA
CTTATTTCCGGAAACCATACCCGAACTCGGATTTTCGGCCGATACGCATATAACGAAACCCCT
CGTTTTCCGACCTAACGCAAAACCGCAACCATACCGCCGCTCGCCTTACGCCCTCGGTA
CGGTTTTCCCTCCGTTCCGCATACCCCTTCAACCAAAACATACATATTAACCTTCGGTATC
TTTCCACAGCTTAAACTTAACTTATTTCCGAAACCGTAAACCGAACCTGATTTTCGACC
CGCGGCCATACGAAACCCCTCGTTTTCCGCCAGCAGCAAAACCGCAACCGCCCTCGCC
CGCTCGTCCGCTACCTCGGTATCGTTTTCCCTCCGTTCCACCGCTACCCCTTCAACCAAAAC
TTACACATTAACCTTCGGTATCTTCTACAGCGCTTAACTTAACTTTTC
>comp3_c0_seq1 len=242 ~FPKM=119.7 2 path=[0]
TTTTTTTTTTTCAAGCAGAAAGCGGCATACGAGCTCTTCGATCTACAAAAAATAAAT
ATTTTTTCGGACCTTCGCGCTTAAACCTTCGTCATATCGCAATCTAAATTTATTTTT
AAACAAATTAATATACGAAAACTACTGTTAAAAACCTTCGGCTTCGGAATACTCTCAA
AAACATAATTTAAACAATATTTCTGAAATATAATACATAAACAGATACCTTCGGACTTAA
TC

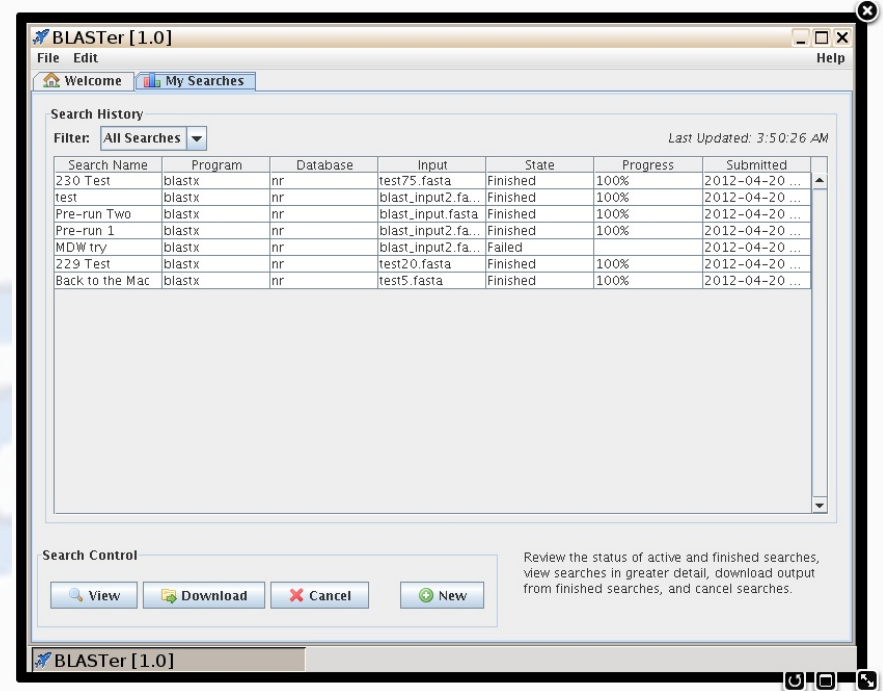
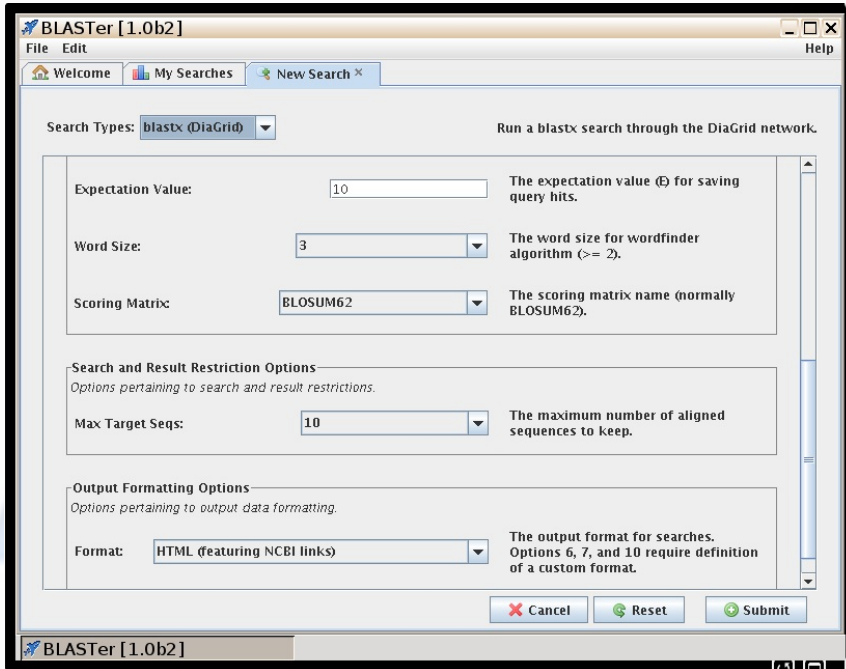
>comp5_c0_seq1 len=1706 ~FPKM=209.9 path=[0]
TTTTTTTTTTTTTCAAGCAGAAAGCGGCATACGAGCTCTTCGCGATCTTCAACAAAAACA
TATCCGAAATTCCTTACCTCTGCTCGCATCGATCAATCGTAAATTCACCTGCTGAA
TAGGAACCTGCTTTTCTCGTACTCTCAGAAATACCAAAATCAGAAATACAAAAA
TCCGACCGCTCGCGACTTCTTCTCATCACACCTTCGGCATATACGCTATATCAACATA
CGGTCGATTTAAAAATACGAACGCGACGTACCAGCATACATAAACCATATATTCGAC
GAACCATCGTTCGACAGTCAAAACCTACGTCGATATCGTAAATCAAAACAAAAA
CCCATCAAAACTTAAAAACGTCAGAAAAATCAATAAATATCTCCGACTTCGAAACCG
TTCATCTCACGCTTCGGCGAAAAAATCTGCTCTGATCCGCTCTGAAACCAACCGGAA
TACTTCACTTAAACCCCTGAAACCGAAAAACCCCTGAAACCTGAAACCCCTCTCACG
AACGGCGCCATCTTAATACCCCGCTCGCGAAAAACCCCTCTTAATCTACGTCGCGCT
ACCATCAGATAATCAGCGCGCGGATCGTAAATCGAAAAACGAAAAAATAACACTG
CTCGTCGAGAAACCGATCTACTCTCATAGTAAATATCTGCGAAACCAATCCGCTAC
CCACAAATTCAGAACTGCTGACGCGGTAATCTGACGCGCGGAAATACGACACTAC
TTCGAATCTCATCCGTAATCTAATATCATCTTCAACCTGAAAAAATCATCCAGTAC
CGAAAAACCTCGGATAAAATACAAAAATAAAGCAAAATCAAAACGAAACAAATCTCG
TTCCGCTCGGAAACCATCAATCCAGATCTTAACGAACTCTGTAACCTGAATAAATC
GACACCGACTCCCAACAGCTCGCATCAACCGGAACTCTGAACCATATTTTTTCGAGGA
TCGCTATAAAAAACGGAACCGCGGCGGACTCTCTCATCTCCGCTCCGGAACAAAC
CTACGCTACGTAACGCTCATTTCCGGCGTCCAAATATACCGGAATACGAAACT
CTGATCAACGGAATACGATCGCCATCGAACTAAAAATCCGACGCTCGACGCTCGCGGT
AACTCGCAGCTGCTCATCGAACCAATCAATAAAAACTCCACTCGCCGACCCGAAAAA
```



Results

comp4_c0_seq1	gi 1996716 gb U00792.1	75.56	45	11	0	377	243	1024	9a-12	73.9
comp4_c0_seq1	gi 12026249 gb U004026.1	75.56	45	11	0	377	243	1024	9a-12	73.9
comp4_c0_seq1	gi 1442832 gb U002037.1	75.56	45	11	0	377	243	1024	9a-12	73.9
comp4_c0_seq1	gi 1011297 gb U004939.1	71.11	45	13	0	377	243	1758	9a-12	71.6
comp4_c0_seq1	gi 1011297 ref U011709.1	68.89	45	14	0	377	243	52	9a-12	68.7
comp4_c0_seq1	gi 1011297 ref U011709.1	66.67	45	15	0	377	243	449	9a-12	66.6
comp4_c0_seq1	gi 1154274 gb U027943.1 U049020.9	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 1996716 gb U00792.1	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 1996716 gb U00792.1	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 1996716 gb U00792.1	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 12026249 gb U004026.1	75.56	45	11	0	377	243	1024	9a-12	73.9
comp4_c0_seq1	gi 1442832 gb U002037.1	75.56	45	11	0	377	243	1024	9a-12	73.9
comp4_c0_seq1	gi 1011297 gb U004939.1	71.11	45	13	0	377	243	1758	9a-12	71.6
comp4_c0_seq1	gi 1011297 ref U011709.1	68.89	45	14	0	377	243	52	9a-12	68.7
comp4_c0_seq1	gi 1011297 ref U011709.1	66.67	45	15	0	377	243	449	9a-12	66.6
comp4_c0_seq1	gi 1154274 gb U027943.1 U049020.9	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 1996716 gb U00792.1	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 1996716 gb U00792.1	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 1996716 gb U00792.1	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 12026249 gb U004026.1	75.56	45	11	0	377	243	1024	9a-12	73.9
comp4_c0_seq1	gi 1442832 gb U002037.1	75.56	45	11	0	377	243	1024	9a-12	73.9
comp4_c0_seq1	gi 1011297 gb U004939.1	71.11	45	13	0	377	243	1758	9a-12	71.6
comp4_c0_seq1	gi 1011297 ref U011709.1	68.89	45	14	0	377	243	52	9a-12	68.7
comp4_c0_seq1	gi 1011297 ref U011709.1	66.67	45	15	0	377	243	449	9a-12	66.6
comp4_c0_seq1	gi 1154274 gb U027943.1 U049020.9	51.16	43	21	0	377	249	1006	9a-15	53.5
comp4_c0_seq1	gi 1996716 gb U00792.1	51.16	43	21	0	377	249	1006	9a-15	53.5





BLASTer [1.0]

File Edit Help

Welcome My Searches test x

test *blastx (DiaGrid)* Running... 9% (12 of 129) Cancel

Log & Statistics
Comments & Parameters
Results

Output Files

- params.txt (365)
- 1.output (9.63 KB)
- 2.output (8.74 KB)
- 11.output (20.33 KB)
- 12.output (22.22 KB)
- 14.output (21.85 KB)
- 17.output (22.1 KB)
- 20.output (12.48 KB)
- 21.output (11.19 KB)
- 23.output (8.55 KB)
- 24.output (11.34 KB)
- 27.output (18.98 KB)
- 28.output (19.69 KB)

File Contents (1.output)

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
14,986,402 sequences; 5,132,328,328 total letters

Query= comp4_c0_seq1 len=401 ~FPKM=247.6
path=[0]

Length=401

Score E

Sequences producing significant alignments:
(Bits) Value

gb|ABF67921.1| j11 putative pol protein [Zea mays]

Download File

Download Archive

As output files become available, they will appear in the list to the left. Select a file to view its contents if possible or to download the file. An archive of all output files can also be downloaded when the search has finished.

BLASTer [1.0]



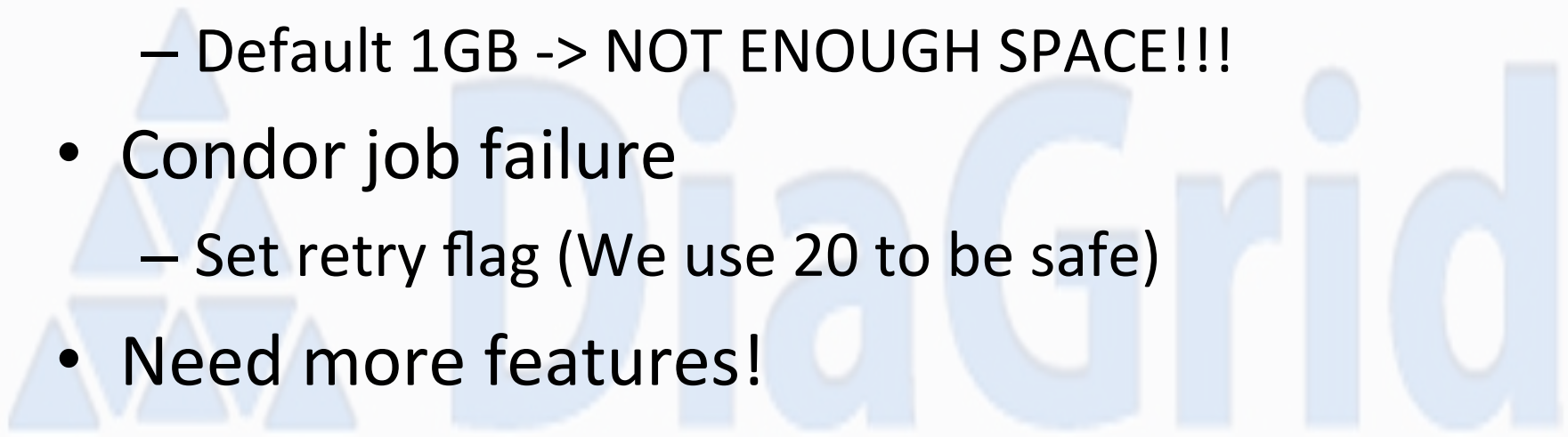
Big Benefits? We think so!

- Rick Westerman
 - Bioinformatics Specialist at the Purdue University Genomics Facility

Sample	Sequences	Bases	N50 length	Cluster wall time	Condor wall time
#1	43 K	16 M	439	5:20	2:40
#2	40 K	15 M	412	5:30	2:30
#3	105 K	49 M	479	7:15	3:30
#4	145 K	72 M	509	8:20	3:20

Development Hurdles

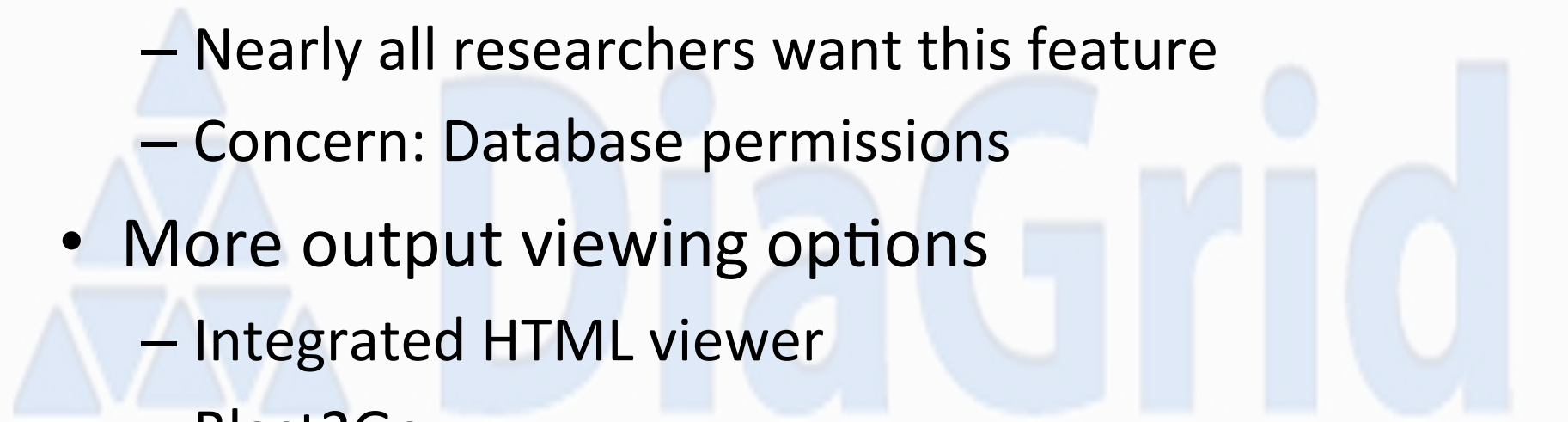
- DiaGrid disk quota per user
 - Default 1GB -> NOT ENOUGH SPACE!!!
- Condor job failure
 - Set retry flag (We use 20 to be safe)
- Need more features!



To the Future!



- Custom Databases
 - Nearly all researchers want this feature
 - Concern: Database permissions
- More output viewing options
 - Integrated HTML viewer
 - Blast2Go
- Better file management



- R (programming language) statistical computing
 - Landscape Ecology & Biodiversity Department
- Cryo-Electron Microscopy Tools (Cryo-EM)
 - Single particle reconstruction (EMAN2 and similar tools)
 - Department of Biological Sciences

Questions?

