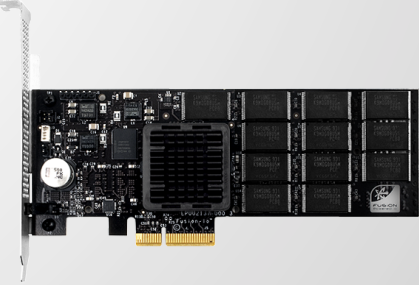
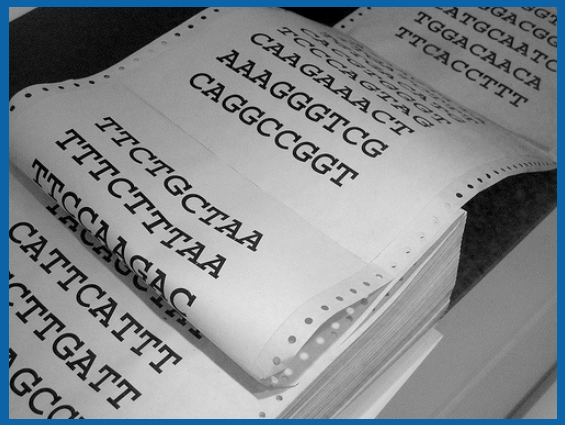
Metagenomic search using new data-intensive computing tools

Jonathan Allen, PhD; LLNL Biosecurity Program

Sasha Ames, Shea Gardner, Tom Slezak, Maya Gokhale, Roger Pearce



Lawrence Livermore National Laboratory



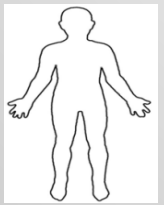
LLNL-PRES-663471

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Metagenomics has the potential to help detect and characterize novel/emerging pathogens

Start by collecting data with *no* prior bias on the sample's contents!

Presents an open computational challenge for fast and accurate database search



Query sample

Sequencing assay

50- 700+ base fragments

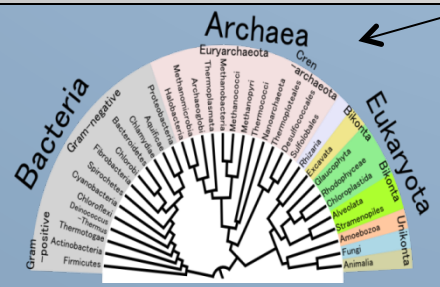
Read1) *attttcaataatactgatactatatac*

Read2) *gttactcaatagtactgatacgtata*

…

ReadN) *ctaactcaatagtactgatacgtat*

Find match against an expanded reference database



Genome Sequence Database

Virulence/ Antibiotic resistance database

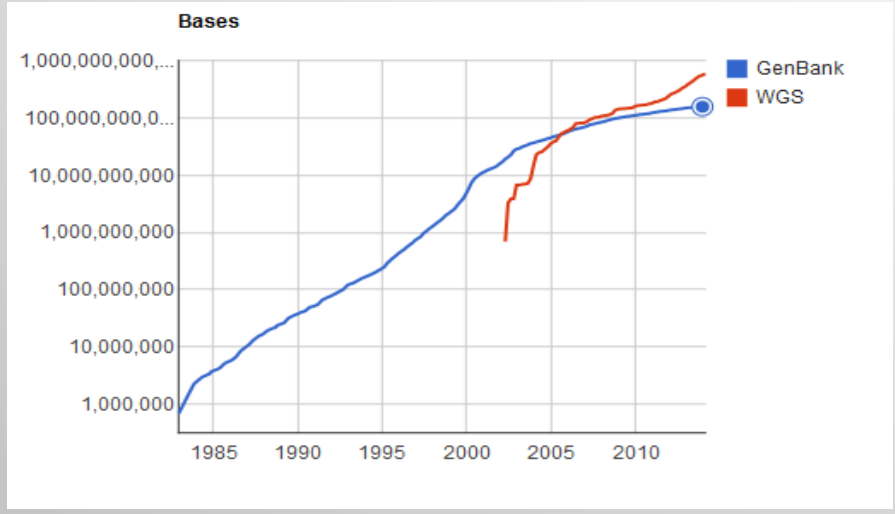
Reference metagenomes

Potential to provide detailed information on difficult to analyze samples:

Microbial species/strain identification, gene function, characterizing novel and engineered threats.

Traditional database search does not scale with increasing sequencer output or meet accuracy requirements

Reference database doubling every 18 months



Sequencer output is increasing

- 1 billion bases (2006)
- 600 billion bases (Now)
- Runtime on a 64,000 CPU HPC
  - 8 hours in 2006
  - 75 days – Now

Single sequencer output!

"One sequencer one compute cluster" would be difficult if not impossible to do with a global distributed network of sequencers

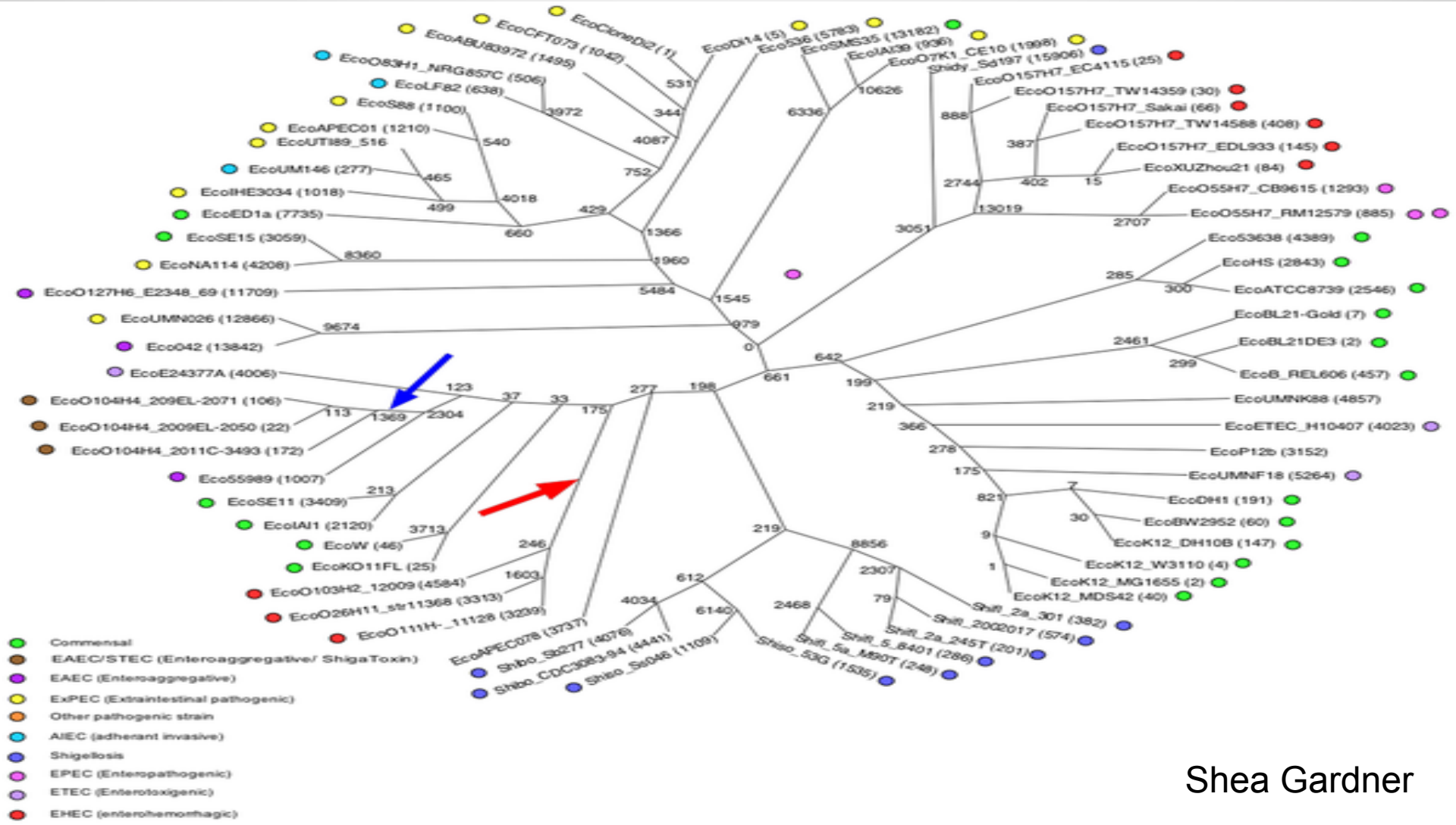
Current methods struggle to detect low abundance pathogens

Current state of the art

1. Reduce query size: use metagenomic '*assembly*' prior to database search (*Pignatelli 2011, Qin 2010*)
   - Still computationally costly, many reads may not assemble, some mis-assembly
2. Reduce database size: use representative collection of sequences, or store the "most information" portions of the genome
   - Information content reduction
   - Scalable but reduced accuracy, cannot tag all reads

Identifying population level genomic variation is important for microbial identification

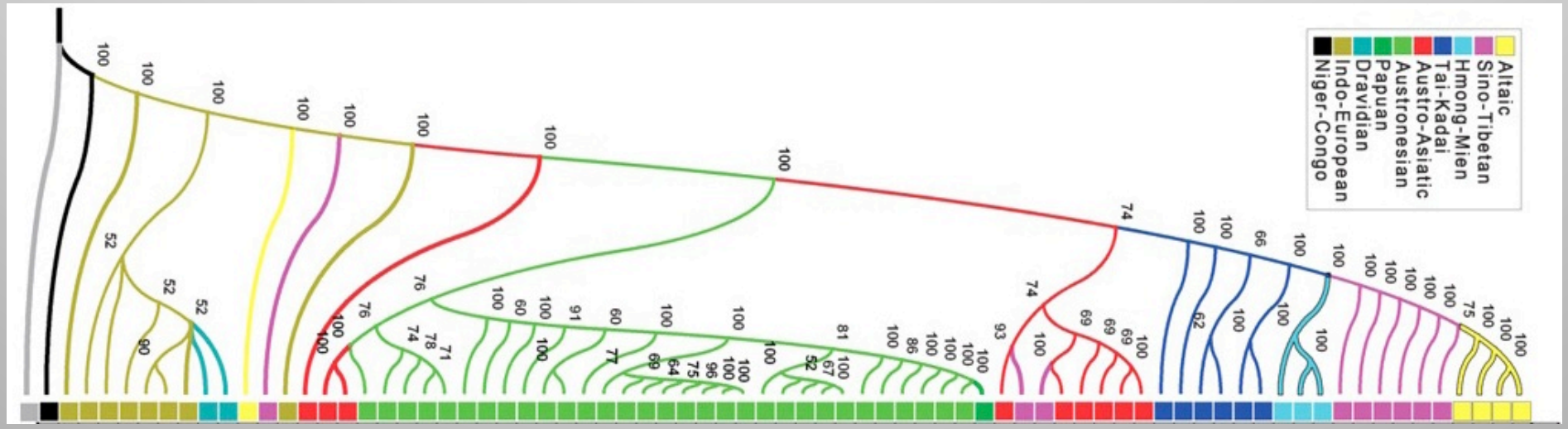
Sub populations of microbes reflect high level of genetic diversity within a species. Each new strain can add up to ~10,000 new nucleotide changes



Shea Gardner

Identifying population level genomic variation is important for host identification

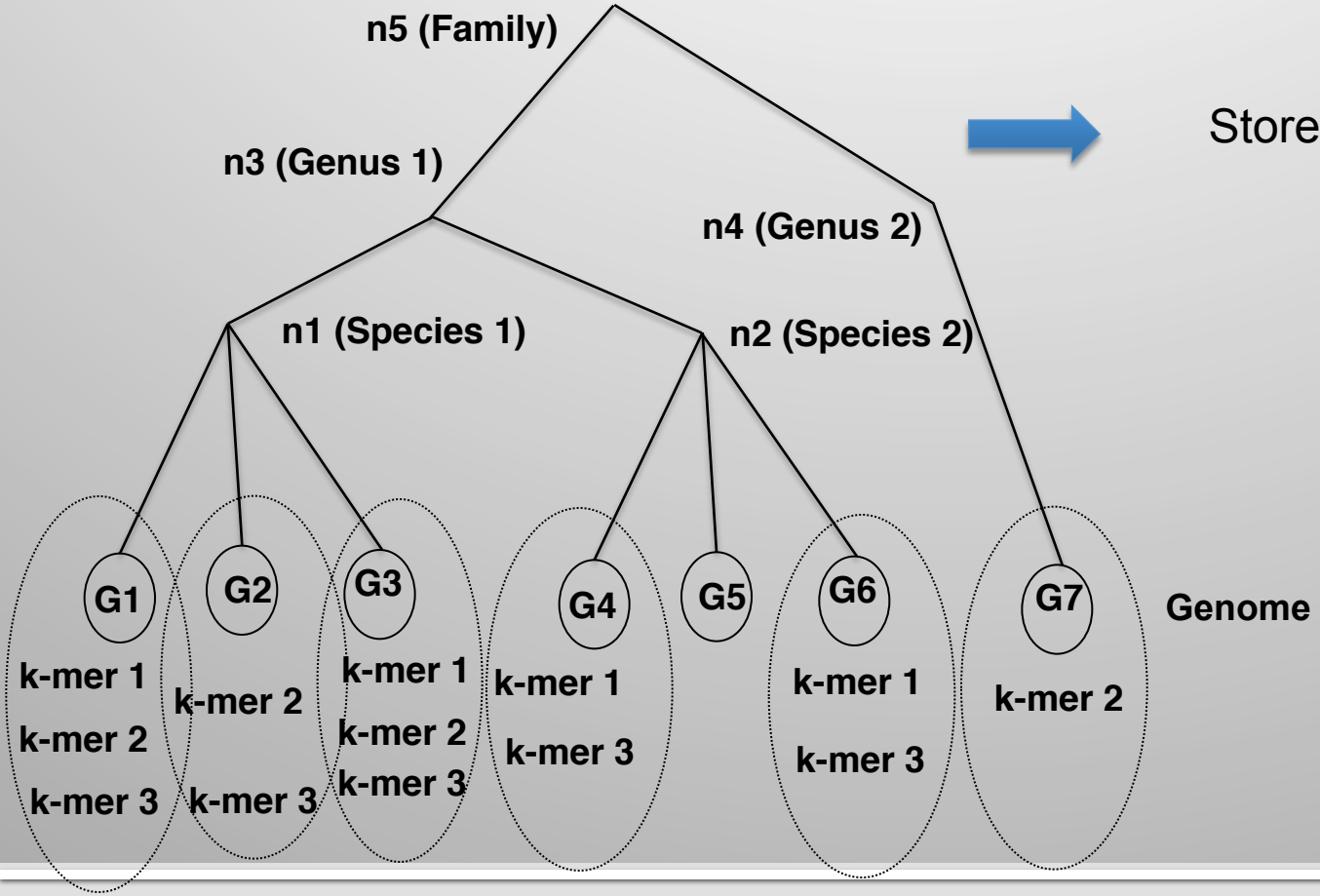
3.6 million nucleotide changes (SNPs) per human & 1.38 million short insertion/deletions (on average)
(1,092 humans - 1000 human genomes project consortium Nature 2012)



(HUGO Consortium, Science 2009)

k-mers store shared genetic relationships between sequenced organisms for rapid information retrieval

k-mer is a short genetic sequence of length k (18-20)
Taxonomy tree captures relationships between k-mers



Store in searchable index

Query read:
k-mer 1, k-mer 2, k-mer 3

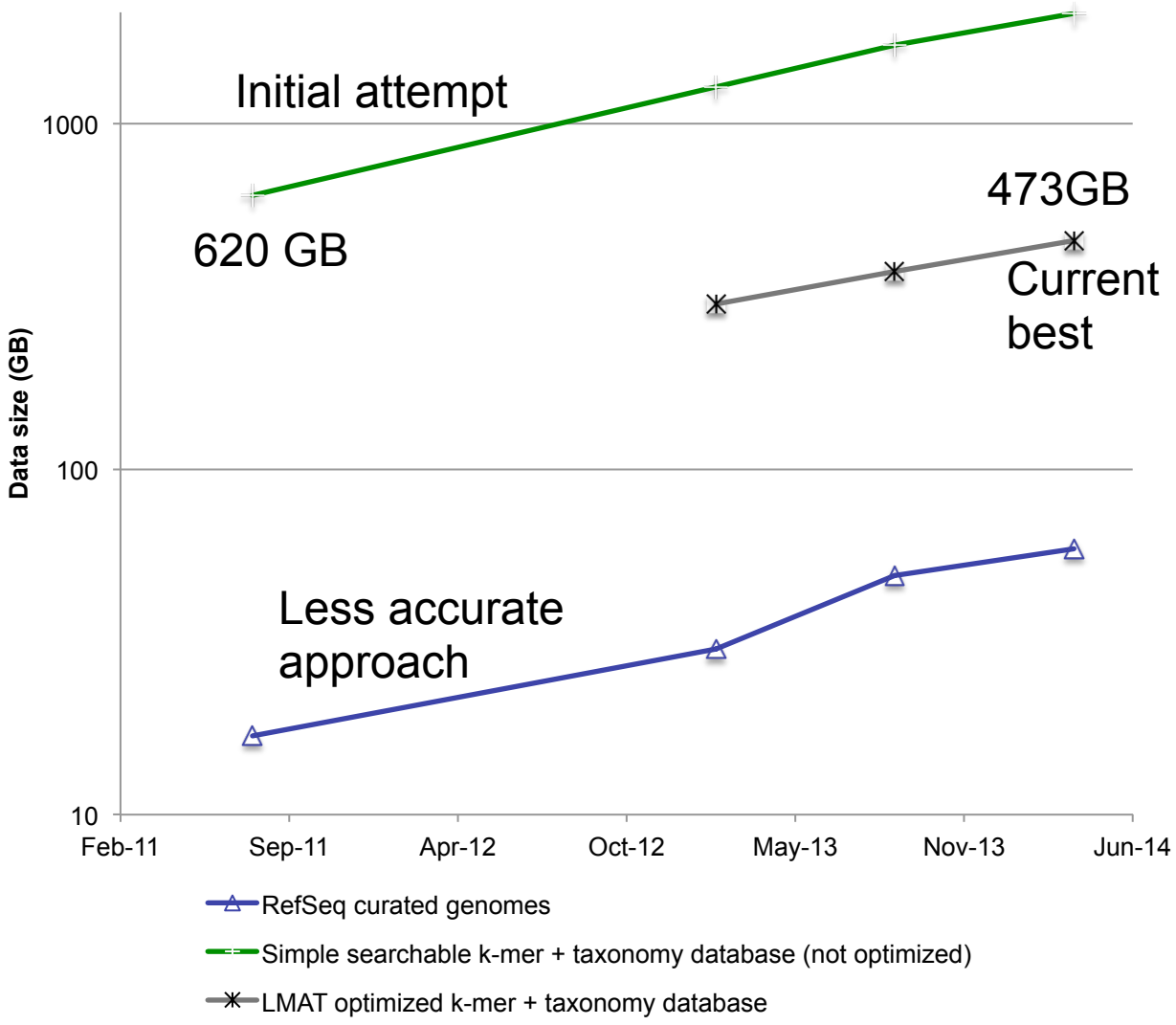
A scoring scheme finds the most rank specific taxonomic label that can be assigned to a sequencer read

Searchable k-mer database
k-mer 1: G1,G3,G4,G6,n1,n2,n3
k-mer 2: G1,G2,G3 ,n1, G7,n5
K-mer 3: G1,G2,G3 ,n1,G4,G6,n2,n3

k-mer	G1	G2	G3	G4	G6	G7	n1	n2	n3	n5
k-mer 1	1	0	1	1	1	0	1	1	1	1
k-mer 2	1	1	1	0	0	1	1	0	0	1
k-mer 3	1	1	1	1	1	0	1	1	1	1
P_j	1.0 (3/3)	0.67 (2/3)	1.0 (3/3)	0.67 (2/3)	0.67 (2/3)	0.33 (1/3)	1.0 (3/3)	0.67 (2/3)	0.67 (2/3)	1.0 (3/3)

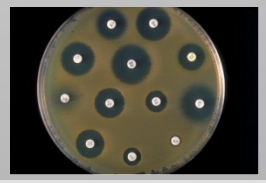
(G1,1.0),(G3,1.0),(n1,1.0),(n5,1.0),(G2,0.67),(G4,0.67),(n3,0.67),(n2,0.67)
(G7,0.33)

Large memory requirements have prevented this approach from being implemented by other groups



All complete and draft genomes:
viruses, bacteria, archaea,
protozoa, fungi, human, and mitochondrial DNA of larger eukaryotes
+ artificial sequence

12,632 species
116+ gigabases of searchable genomic data
~3 times larger than the closest searchable DB



Population Level for the 1st time!

Inclusion of draft genomes dramatically improves read binning sensitivity

Database	All (%)	RefSeq Only (%)
Weak Match	14.81	42.34
No Match	0.05	2.76
DB Size	116 Gbase	9.13 Gbase

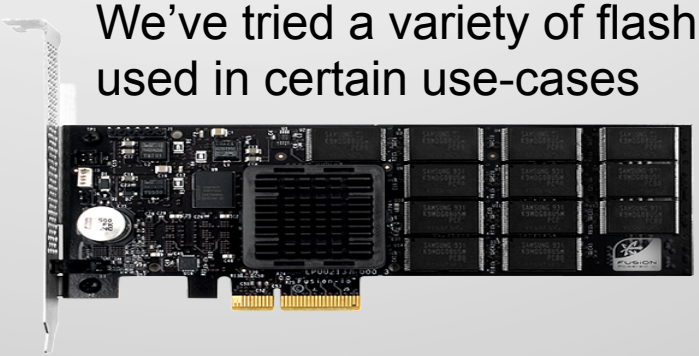
Percentage of reads given no taxonomic assignment (No Match) or low confidence assignment (Weak Match) – 131 HMP samples

Search algorithms designed to use NVRAM could give competitive performance to traditional DRAM only platforms

$70K DRAM compute node could be replaced with 3 $20K DRAM+NVRAM compute nodes

Use flash drive (NVRAM) as extended memory by memory mapping database file

We've tried a variety of flash drives, each can be used in certain use-cases



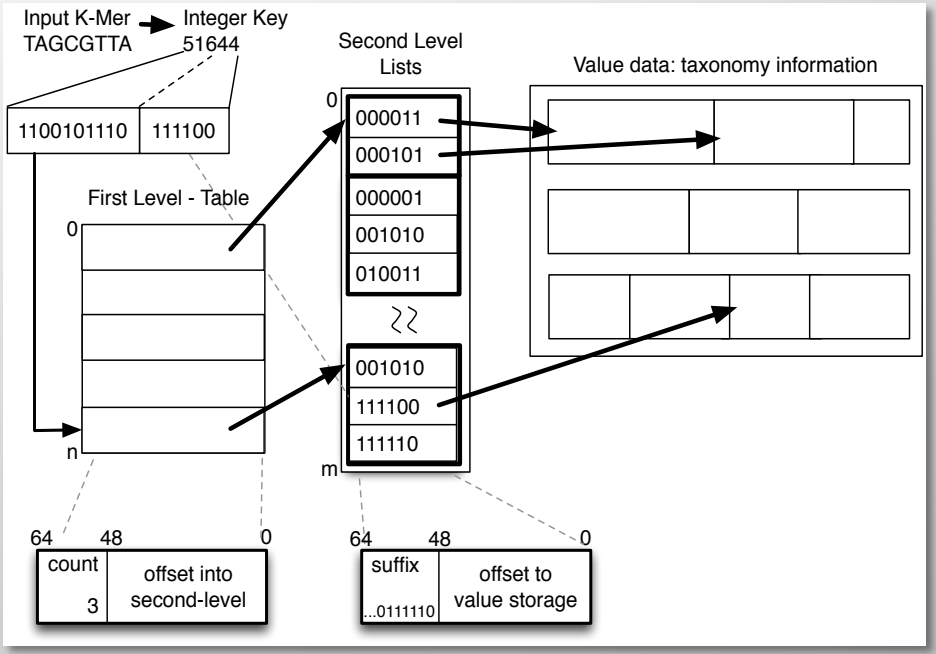
High end PCIe SSD

Mid-tier PCIe SSD

Low cost SATA drives

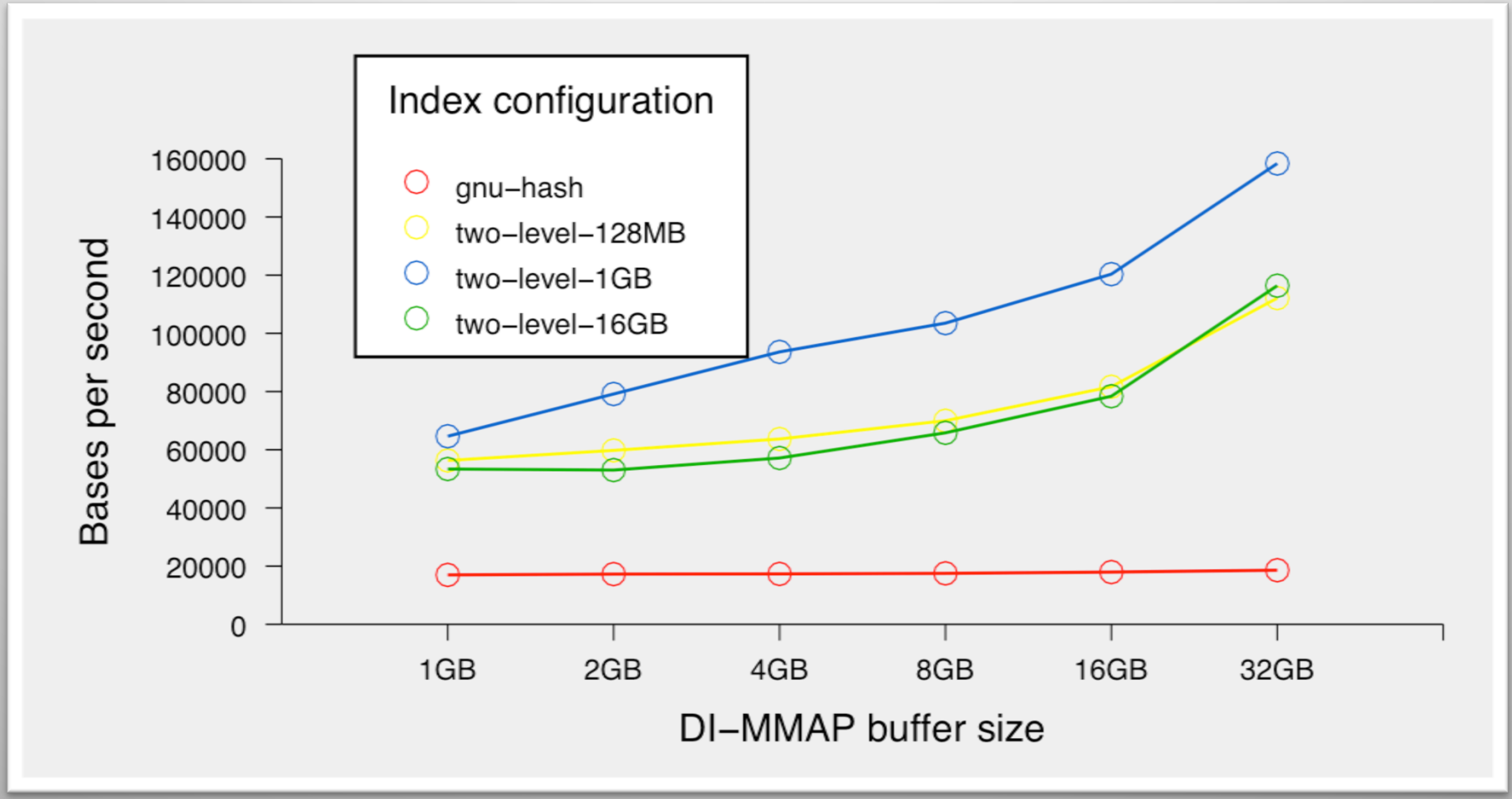
Two-level k-mer index used to access NVRAM

- Goal: increase locality for use with page oriented access
- Split integer keys
- First level: map prefix to “chunk” in second level
- Binary search second level lists
- Tradeoff lies in how to split key, configure level sizes.



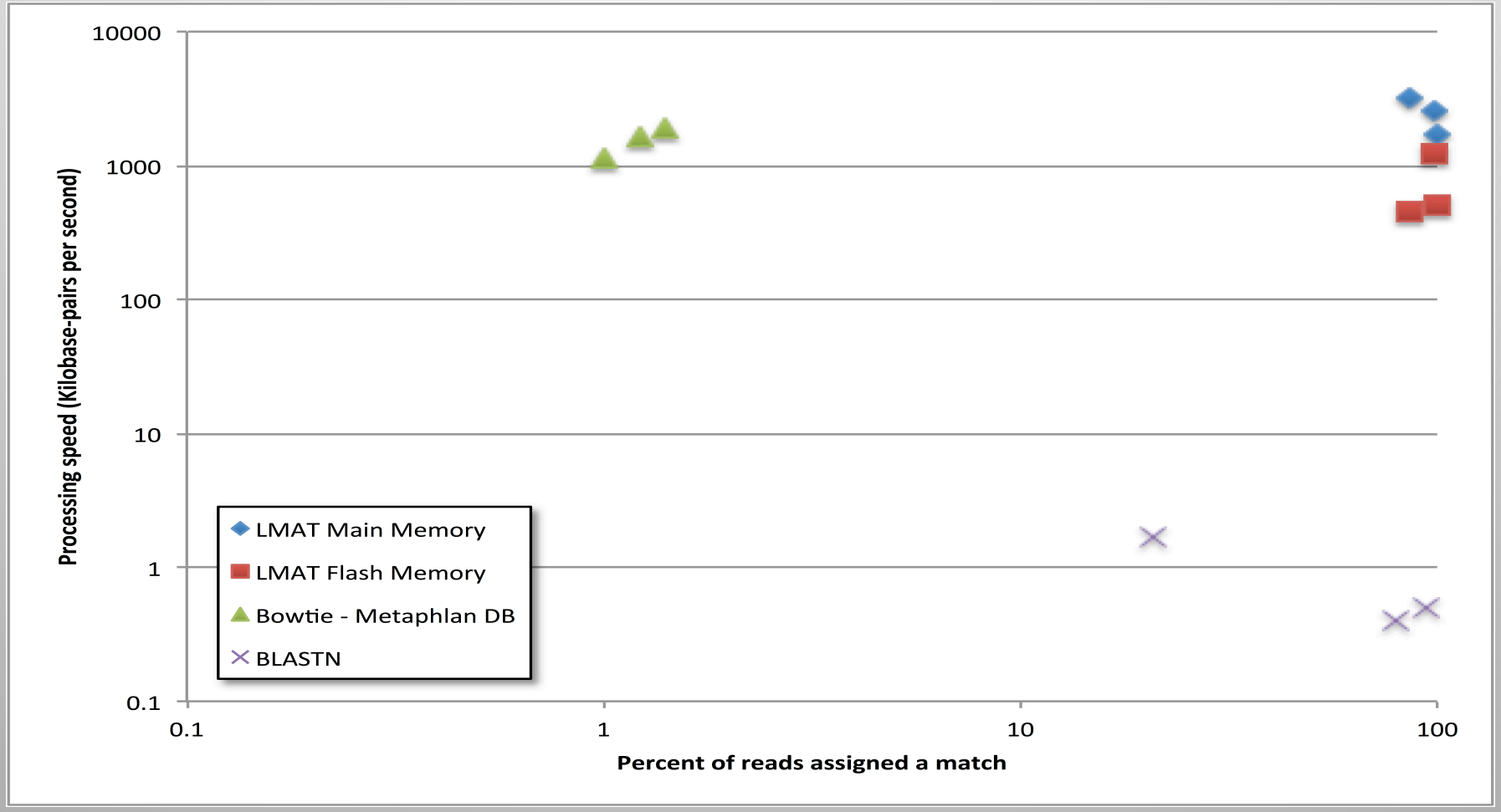
Use NVRAM to improve scaling and cost

Demonstrated speed up with NVRAM using a 2-level index over a traditional hash table

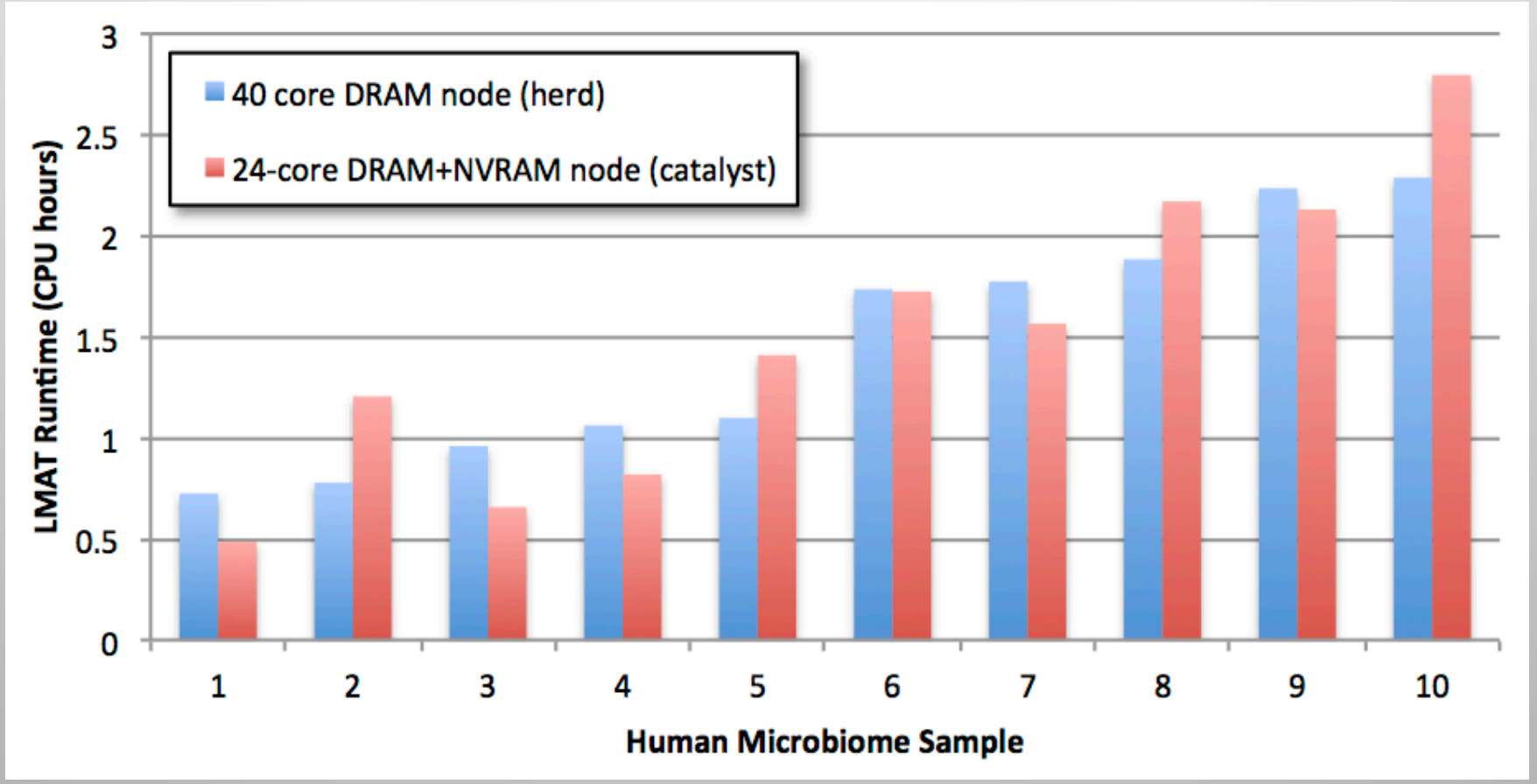


DI-MMAP is memory mapping kernel module (Brian Van Essen)

NVRAM gives competitive performance to traditional platforms with only 64 GB DRAM



Increasing to 128 DRAM + NVRAM is showing comparable performance to DRAM only

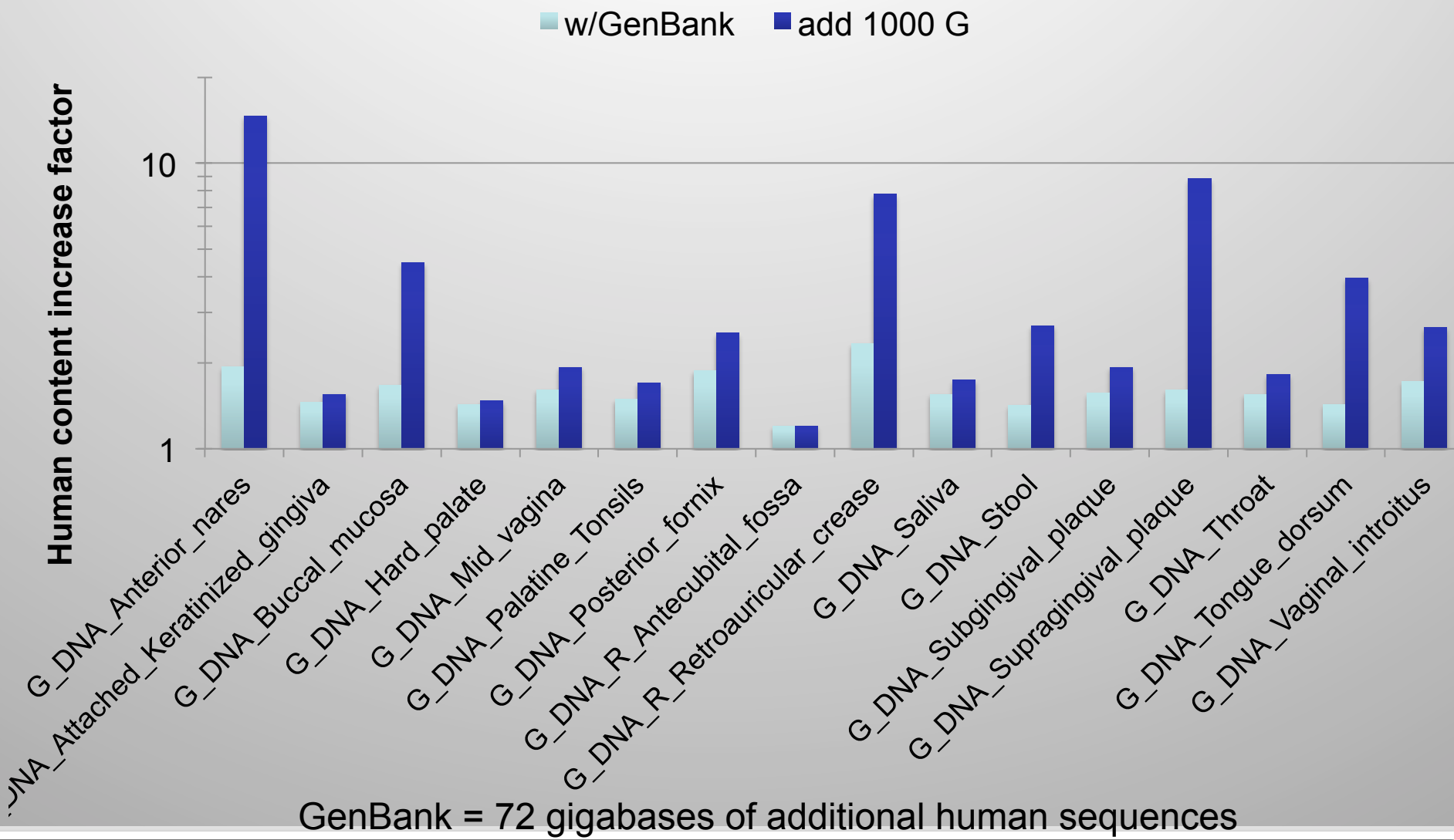


Access to NVRAM through Catalyst allows us to scale up our analysis to an unprecedented scale

8 times more memory per core than typical (39 GB); 928 GB of memory per node, (800 GB NVRAM, 128 DRAM)
304 nodes with 24-core Intel ivy bridge processors
~194x speed up over our single 40-core compute node server

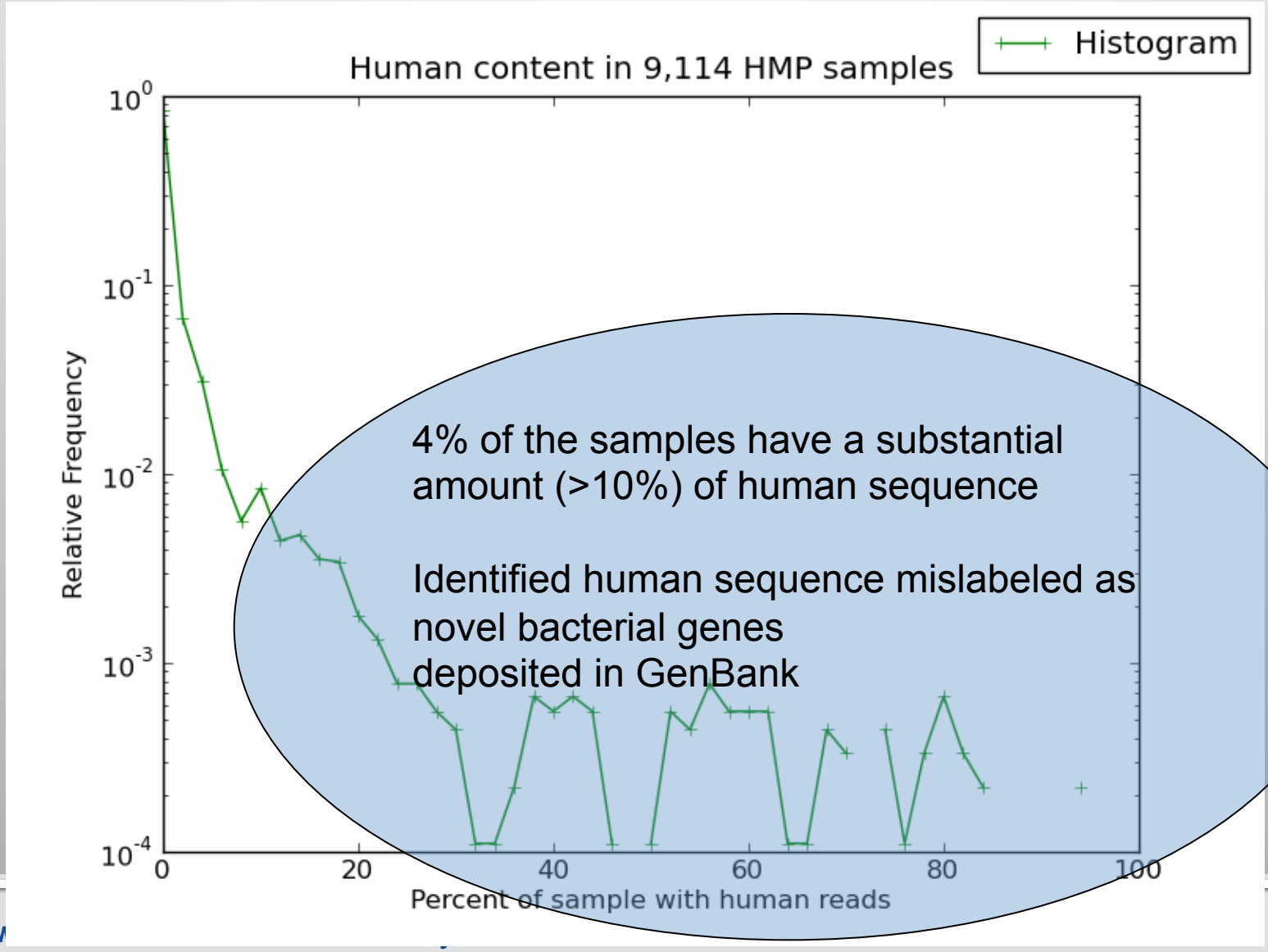
Collection	Type	Size	Run time	Result
1000 Genome Project (2,646 people)	Human (mostly)	90 terabases (200 compressed terabytes)	6 days	Added 8 million new genetic variants (208 million from GnBk)
Human Microbiome (9,114 samples)	Microbes (mostly)	18 terabases	26 hours	Discovering new human sequence

Enrichment for human DNA is not randomly distributed across body sites and cannot be easily detected without human population data



GenBank = 72 gigabases of additional human sequences

Samples thought to contain only microbial DNA have substantial amounts of human DNA which can confound the pathogen identification process



1000 human genomes project provides a valuable "negative" control

Result: sensitivity is a both a blessing and a curse!

- Presents new challenges for medical diagnostics
- Sequencer "carryover" leads to the potential presence of pathogens
  - 10^-4 was the minimum abundance in 1000 humans
- See viruses that infect cell culture used to preserve human sample and any DNA

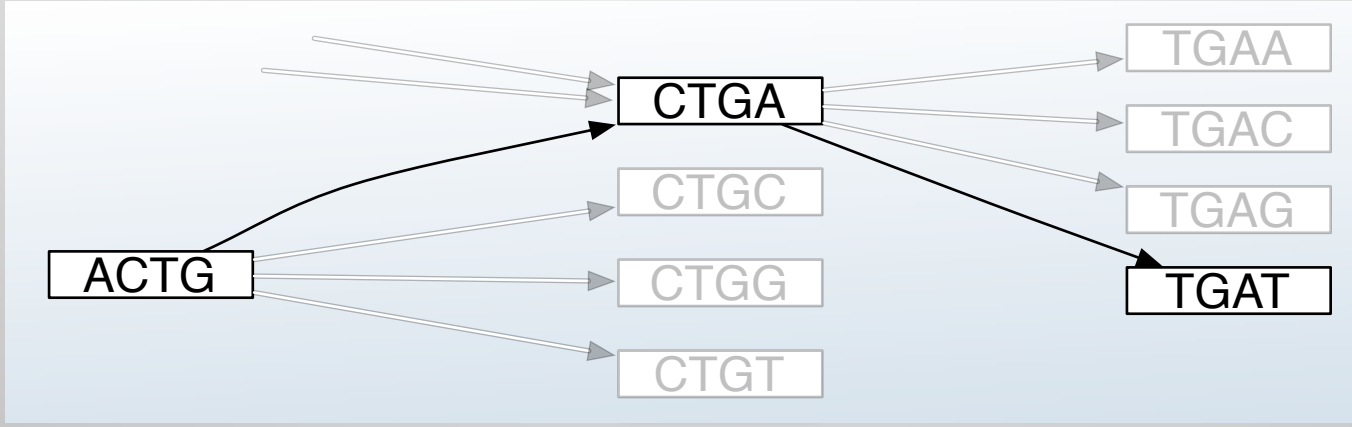
More accurate detection of pathogens in a complex sample is now possible and helping to advance the field of sequence based diagnostics

Ongoing challenges

- Improving runtimes (without sacrificing accuracy) is still highly desirable!!
  - Analysis of bench top sequencing can still take 2-3 hours;
  - Our reduced size database addresses this problem, but can we do better?
- How accurate are abundance estimates using k-mer distributions, compared to a traditional read mapping approach? Can we get fast accurate abundance measurements?
- Adding the novel read characterization component using existing tools.
  - We've reduced the 18 TB of HMP data to a few hundred megabases of unlabeled reads; can we find anything interesting hidden in there?
- Can we use positional information to add more large genomes and identify important functional features?

De Bruijn graph is a powerful tool for efficient storage and retrieval of shared relationships in genomic data

ACTGAT (Bold Path In Graph)

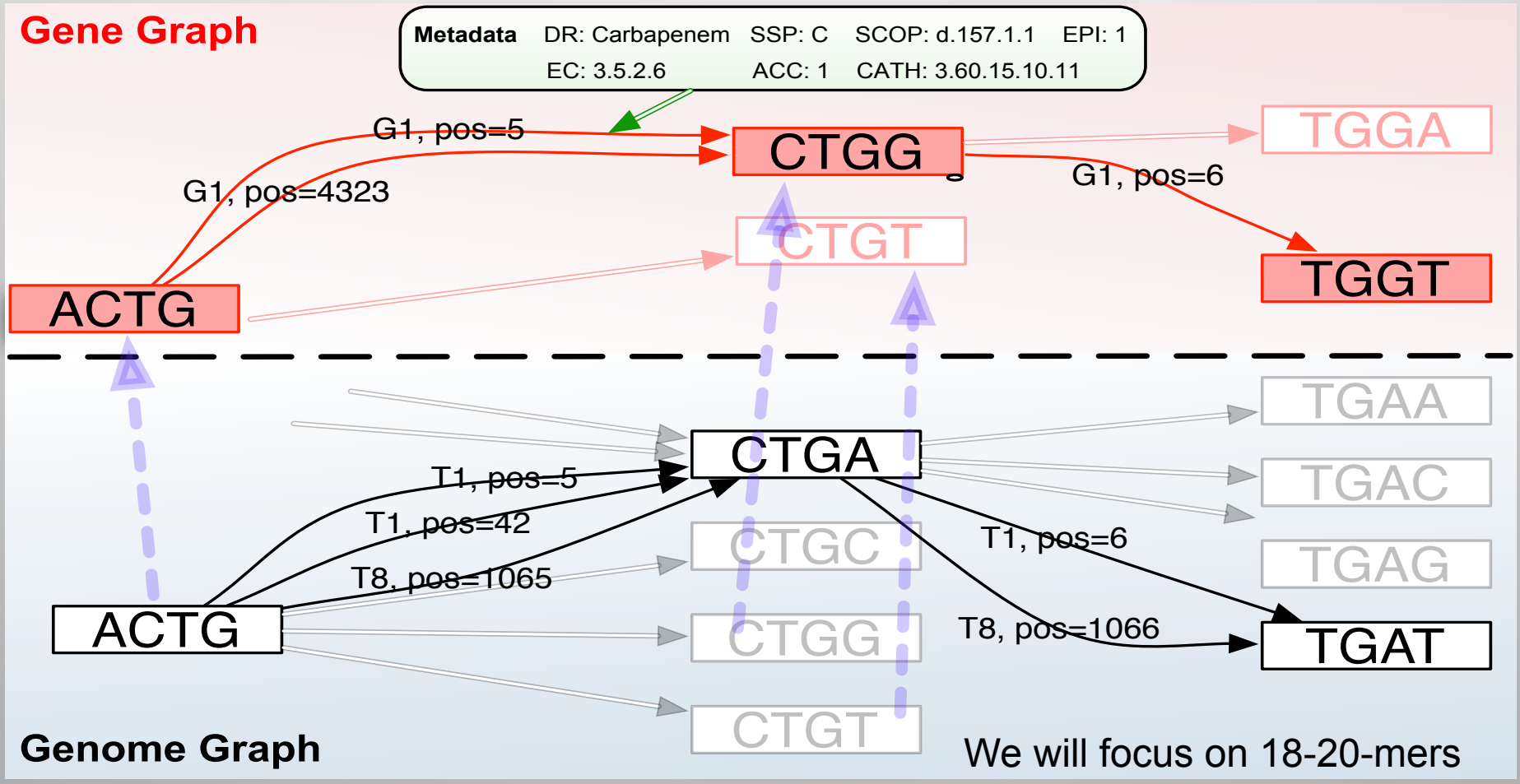


Sequence assembly – exploit ability to process large amount of sequencer data.
Multiple sequence alignment – exploit ability to capture re-ordering of sequences.

Applications are memory intensive and have been traditionally limited by availability of DRAM.

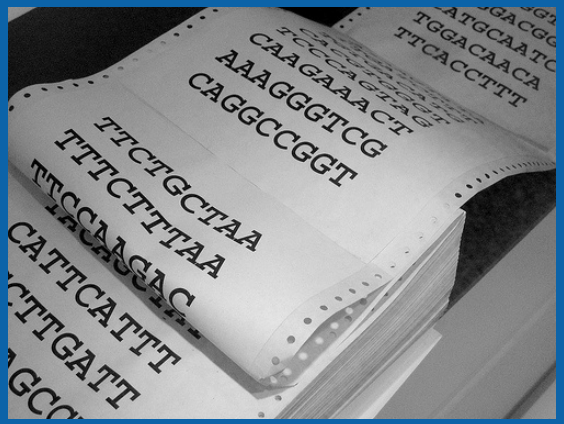
Marcus et al., 2014 (Schatz Lab)

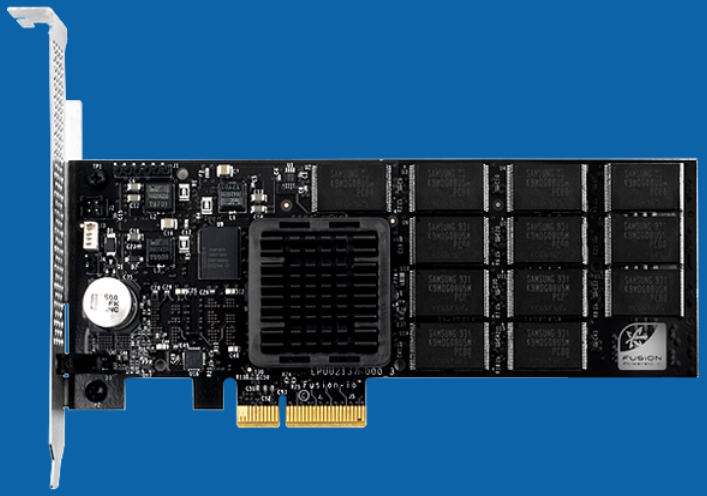
Using a multi-scale de Bruijn graph as a new flexible graph structure for genomic search



Questions?

LMAT is open source, software and databases available for download: http://sourceforge.net/projects/lmat/





Lawrence Livermore National Laboratory