LA-UR- 10-08068

Approved for public release;
distribution is unlimited.

Title:	Shotgun Metagenomic Data Streams: Surfing Without Fear
Author(s):	Joel Berendzen
Intended for:	4th National Bio-Threat Conference, New Orleans, LA, Dec 9th

Form 836 (7/06)

Shotgun Metagenomic Data Streams: Surfing without Fear

Joel Berendzen, Los Alamos National Laboratory

Timely information about bio-threat prevalence, consequence, propagation, attribution, and mitigation is needed to support decision-making, both routinely and in a crisis. One DNA sequencer can stream 25 Gbp of information per day, but sampling strategies and analysis techniques are needed to turn raw sequencing power into actionable knowledge.

Shotgun metagenomics can enable biosurveillance at the level of a single city, hospital, or airplane. Metagenomics characterizes viruses and bacteria from complex environments such as soil, air filters, or sewage. Unlike targeted-primer-based sequencing, shotgun methods are not blind to sequences that are truly novel, and they can measure absolute prevalence. Shotgun metagenomic sampling can be non-invasive, efficient, and inexpensive while being informative.

We have developed analysis techniques for shotgun metagenomic sequencing that rely upon phylogenetic signature patterns. They work by indexing local sequence patterns in a manner similar to web search engines. Our methods are laptop-fast and favorable scaling properties ensure they will be sustainable as sequencing methods grow. We show examples of application to soil metagenomic samples.

Shotgun Metagenomic Data Streams: Surfing Without Fear

Joel Berendzen

Los Alamos National Laboratory

joelb@lanl.gov

Collaborators & Funding

- Ben McMahon, Nick Hengartner, Judith Cohn (LANL) Metagenomic Data Analysis
- Cheryl Kuske (LANL), Michael Blum (Tulane): Metagenomic Data Collection
- Patrick Chain (LANL): Metagenomic Sequencing
- Carla Kuiken (LANL): Viral Pathogen Short Read Analysis
- Jeanne Fair (LANL): Zoonotic Epidemiology
- Murray Wolinsky (LANL): Sequence Signature Analysis
- Funding: LANL Laboratory-Directed R&D (LDRD), DTRA

Metagenomics: Characterizing Microbial Communities

- **Metagenomics: communities of microbes by sampling DNA or RNA from complex environments such as dirt, air filters, or sewage.**

Oiled marshland, Louisiana

Airliner air filter

Airliner lav service

No need for blood samples or nasal swabs, (usually) not on individuals → few privacy issues.

Shotgun Metagenomics: Look at All the (High-Frequency) Sequences, Not Just a Few

- Shotgun: all the genes present in a sample

The ISME Journal (2009) 3, 1365–1373

www.nature.com/ismej

ORIGINAL ARTICLE

Polymerase chain reaction primers miss half of rRNA microbial diversity

OPEN ACCESS Freely available online PLoS one

Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities

Jack A. Gilbert[1]*, Dawn Field[2], Ying Huang[3], Rob Edwards[4,5], Weizhong Li[3], Paul Gilna[3], Ian Joint[1]

Targeting/amplification usually leads to targeted blindness.
PCR is never quantitative.

Shotgun Metagenomics Can Find Bacterial Bio-Threat Genes in Complex Environments

Metagenomic Diagnosis of Bacterial Infections

Emerging Infectious Diseases,2008

Science
AAAS

A Metagenomic Survey of Microbes in Honey Bee Colony Collapse Disorder
Diana L. Cox-Foster, *et al.*
Science 318, 283 (2007);
DOI: 10.1126/science.1146498

Vol 464| 4 March 2010|doi:10.1038/nature08821 nature

ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

Shotgun Metagenomics Can Find Viral Bio-Threat Genes in Complex Environments

OPEN ACCESS Freely available online

PLoS one

Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach

nature Vol 466|15 July 2010|doi:10.1038/nature09199

ARTICLES

Viruses in the faecal microbiota of monozygotic twins and their mothers

JOURNAL OF VIROLOGY, July 2010, p. 6955–6965 Vol. 84, No. 14
0022-538X/10/$12.00 doi:10.1128/JVI.00501-10

Bat Guano Virome: Predominance of Dietary Viruses from Insects and Plants plus Novel Mammalian Viruses[▽]

UNCLASSIFIED

Antibiotic Resistance is a Another Ecological (Community) Problem That Metagenomics Can Address

- Antibiotic resistance predates human antibiotic use
- Main mechanism of antibiotic resistance is horizontal gene transfer
- Horizontal gene transfer happens *very* frequently
- The gene-transfer agents are virus-like particles taken up by bacteria

Generally Overlooked Fundamentals of Bacterial Genetics and Ecology

Anne O. Summers
Department of Microbiology, University of Georgia, Athens

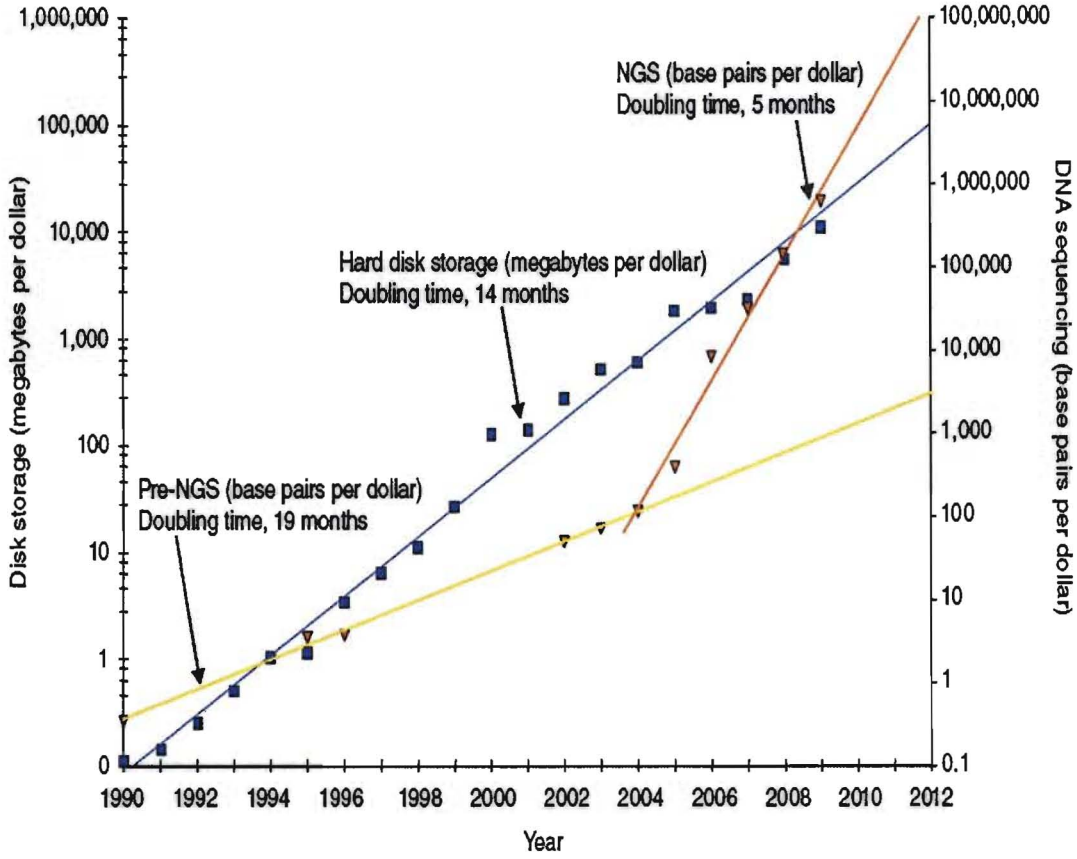
BREVIA

High Frequency of Horizontal Gene Transfer in the Oceans

Lauren D. McDaniel,[1]* Elizabeth Young,[1] Jennifer Delaney,[1] Fabian Ruhnau,[2] Kim B. Ritchie,[3] John H. Paul[1]

Shotgun Metagenomics Can Be Cost-Effective: Sequencing is No Longer the Dominant Cost

- **Late 2010 expendables cost of sequencing one:**
  - Human genome: $5K
  - Bacterium: 30 cents
  - Megabase: 10 cents
  - Virus: 0.3 cents
- **It costs less to sequence 1 DNA base pair than to store 1 MB on disk**
- **Analysis is limiting: "$1,000 genome, but $1,000,000 interpretation"**

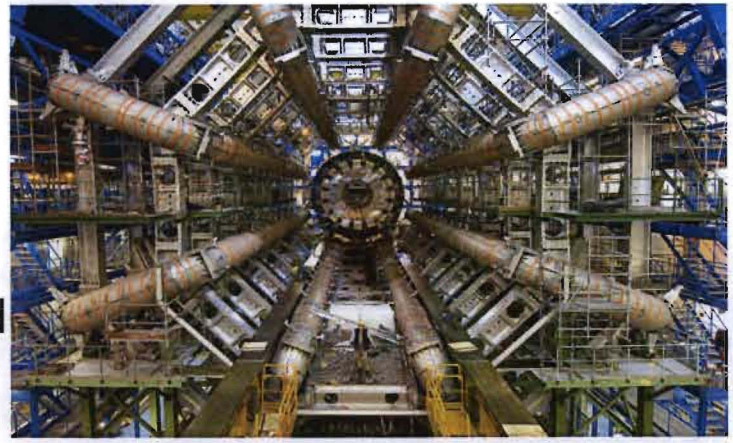


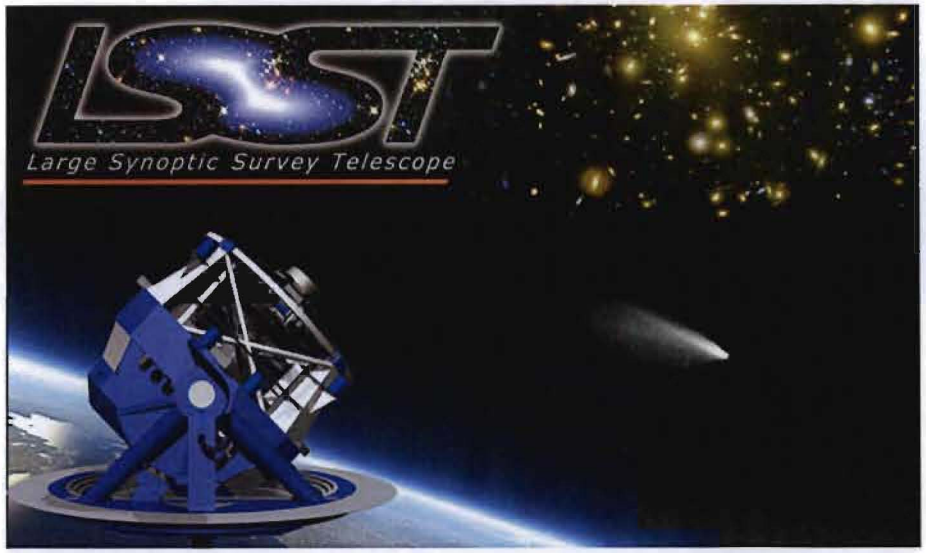
Stein, L.D., *Genome Biol.* **11**, 207 (2010).

Analysis Challenges & Needs for Shotgun Metagenomics Analysis for Biosurveillance

- **Challenges:**
  - Lots (>50 M/sequencer run) of short (75-150 bp) reads
  - Assembly is not the paradigm
    - Reads are usually unassembleable
    - Per-read quantitation is key to prevalence
  - BLAST doesn't work for short reads
- **Needs:**
  - Streaming data analysis, 25 GB/sequencer-day
    - 1 TB/day= LHC, LSST big science
  - Prevalence (including denominator)
  - Similarity of complex data streams to

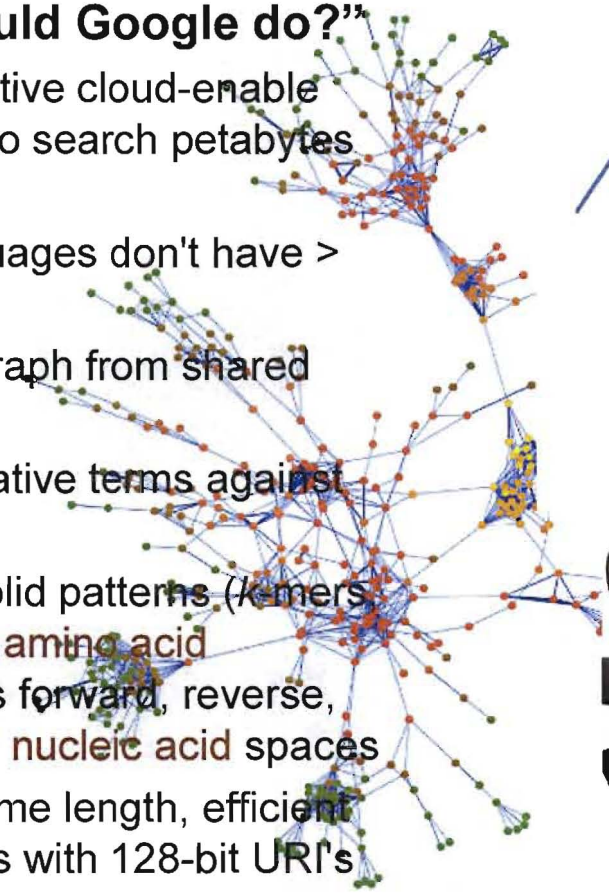
LHC
&
LSST
~ 1TB/d

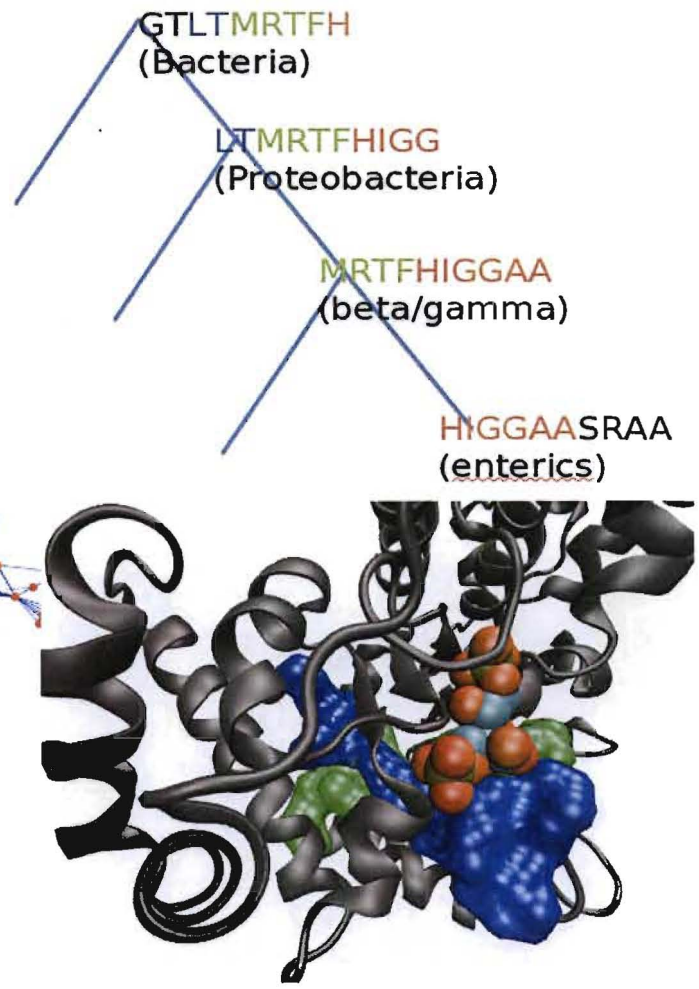




Phylogenetically-Aware Sequence Indexing Methods Offer a Solution

- **Indexing: "What would Google do?"**
  - Fast, scalable, definitive cloud-enable app, solutions exist to search petabytes of data
    - But human languages don't have > 1B words
  - Construct network graph from shared terms in index
  - Search list of informative terms against index
  - Index overlapping solid patterns (*k*-mers, *N*-grams) in 6-frame amino acid translation as well as forward, reverse, reverse-complement nucleic acid spaces
    - Words are all same length, efficient as 64-bit integers with 128-bit URI's

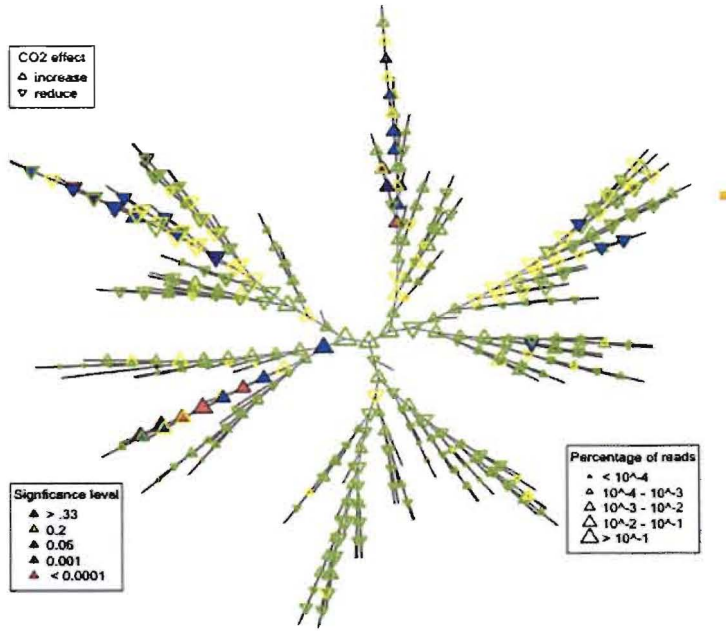


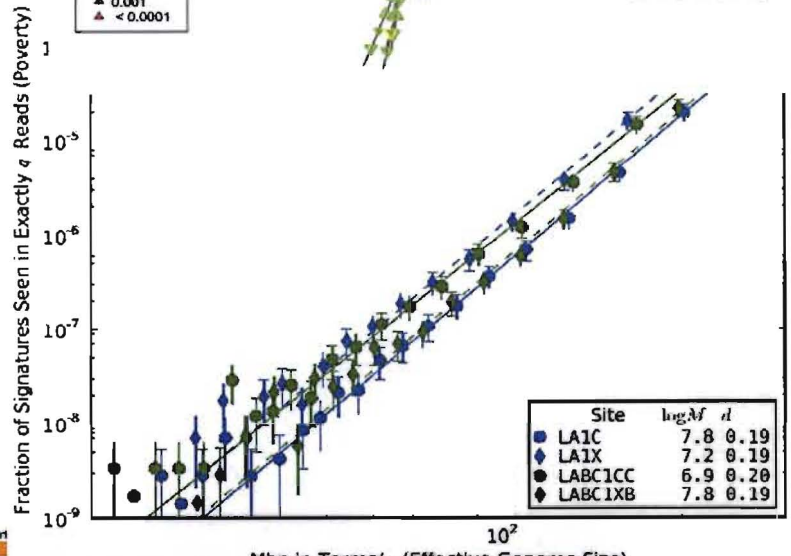


- **Phylogenetic Signatures:**
  - Use minimal distance cutoffs to construct an efficient list of search terms
    - 20:1 reduction at genus level in

Sequedex: Indexing and Searching the Tree of Life

- **Tasks: Index, search, and display phylogeny, function, complexity, and similarity among multiple metagenomic data sets**
  - Phylogeny assigned to interior nodes of tree
  - Can focus on functions of interest (e.g. pathogenicity genes)
  - Complexity, similarity measures are new
  - Time- and geography-dependence
- **Performance: laptop metagenomics**
  - Keeps up with sequencing: <1 CPU-day per next-gen sequencer data set (~x10,000 faster than BLAST)
  - Data-parallel, many cores and servers
  - $O(N \log N)$ scaling





Characterizing phylogeny, complexity, and similarity of soil samples at oil spill-affected sites.

Some Bio-Threat Problems Shotgun Metagenomics + *Sequedex* Can (Uniquely) Address: 1

- **Prevalence: How common is disease X at location Y?**
  - X= flu, dengue, MRSA, hemorragic fevers...
  - Y=airport/airplane, hospital, city, operational site...
  - *Sequedex* can measure the denominator that public health systems and targeted monitoring can't get
  - Prevalence is key input for epidemiological models
- **Novelty: What's different in a stream of sequence data?**
  - Novelty of high-frequency genes is a *Sequedex* strength
  - Something will be new in every sample
  - Need significant differences across space and time
- **Consequence: What impacts can we expect from differences?**
  - *Sequedex* can assign phylogeny, function, pathogenicity to a high % of reads
  - Correlation with history is key. *Ad hoc* solutions aren't as trustworthy as streams with track records.

Importance ~ (prevalence x consequence)

Some Bio-Threat Problems Shotgun Metagenomics + *Sequedex* Can (Uniquely) Address: 2

- **Propagation: How is this spreading?**
  - Breadth of sampling is a unique metagenomics strength
  - e.g., antibiotic resistance moving through the hospital chain (battlefield → military hospitals → civilian hospitals)
  - Time & spatial distributions key to constraining epi models
- **Attribution: Where did this come from?**
  - Metagenomics can measure phylogeny, host/geographic distributions
  - Metagenomics can answer "How likely is this to be an engineered threat?"
- **Mitigation: How much time are we buying, at what cost?**
  - Metagenomics and *Sequedex* can measure prompt effects from applied mitigations such as
    - Closing airports, schools, or businesses
    - Discouraging/prohibiting transport
    - Exterminating disease vectors
    - Changing hospital procedures

Shotgun Metagenomics + *Sequedex* Can Produce Actionable Knowledge at a Sustainable Level of Effort

- **Index-based analysis is fast, scalable, informative**
- **Can give unique knowledge about bio-threat prevalence, novelty, consequence, propagation, attribution, and mitigation**
- **Prevalence outputs can be fed into epi models that can directly inform decision-making**
- **A commitment to data streams now could give timely, relevant knowledge in a crisis**
  - Emergent viral threats (e.g., hemorragic fevers)
  - Evolution of antibiotic resistance (e.g., NDM-1)
  - Spread of seasonal and epidemic influenzas

Sewage diver, Australia