

# NHGRI/NCBI Short-Read Archive: Data Retrieval

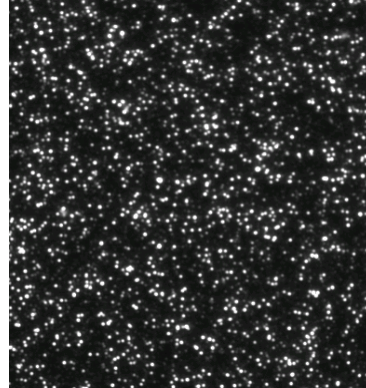


Gabor T. Marth  
Boston College Biology Department  
<http://bioinformatics.bc.edu/marthlab/>

NCBI/NHGRI Short-Read Archive meeting  
Bethesda, MD, July 27, 2007

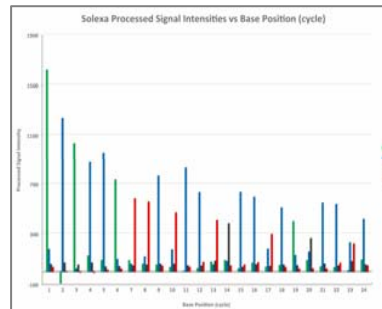
## What level of raw data detail is required for 3rd party downstream applications?

images



## ◀ 3<sup>rd</sup> party image analysis?

processed  
traces (color  
intensities)



## ◀ 3<sup>rd</sup> party base callers?

base sequence +  
quality values

```
>B_TITR_1_1_668_35 TIME: Tue Feb 20 02:26:06 2007
ATATCGGATGACACAATATGGGAGGTTGAC
>B_TITR_1_2_843_403 TIME: Tue Feb 20 02:26:06 2007
TGTAGCTTTTCATGACAATTTTATAGGTGT
```

[illegible]

# What metadata is needed for specific analyses?

## General:

- Machine HW and read processing SW versions
- Organism
- Library construction details
- Genomic DNA, cDNA, bisulfite-treated DNA, etc.?
- Single-end or paired-end reads?

## Alignment / Assembly:

- Attempted read length
- Potential sequence clipping (quality; vector, linker, primer, barcode)
- Reference genome sequence for re-sequencing applications (genome / transcriptome; whole genome, target region)



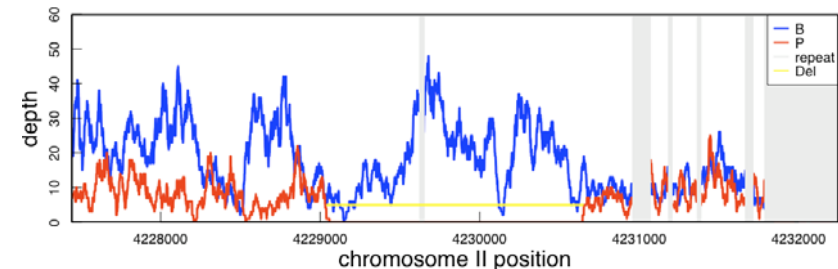
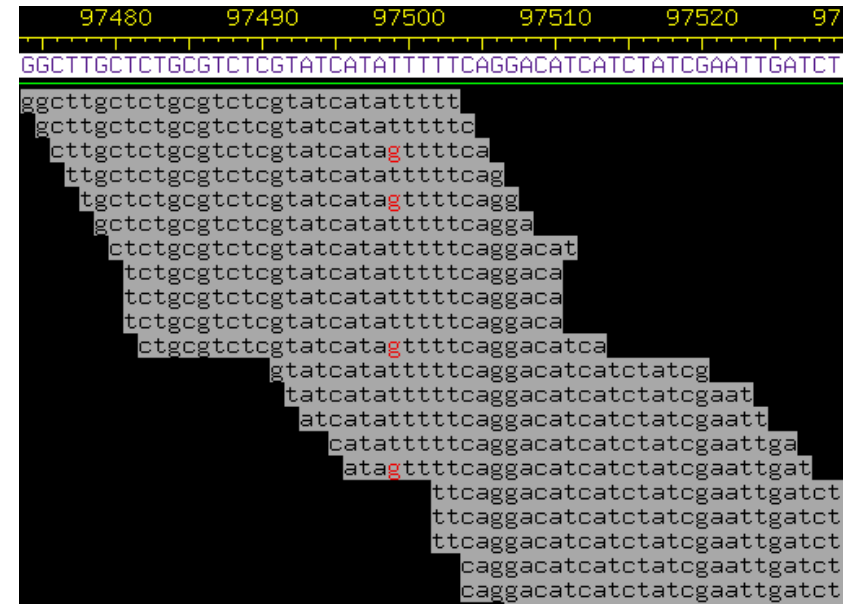
# Metadata (cont'd)

## SNP discovery:

- • Base quality values
- **Traces???**
- Source DNA (diploid genome / PCR vs. clonal; single individual vs. pooled)
- Sample phenotype / disease status (tumor / normal)
- Ethnicity?

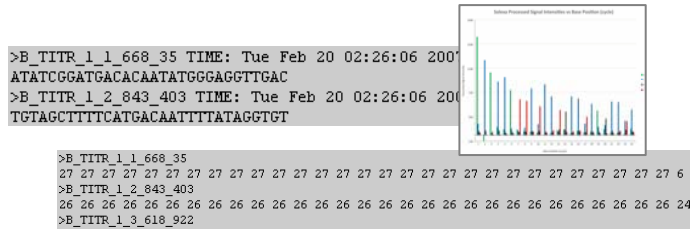
## Structural variation detection:

- • Fragment length range
- Mate-pair relationship
- Ploidy

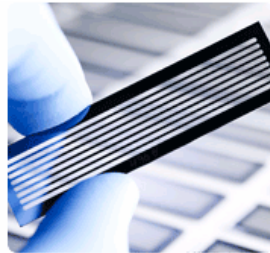


- Most read attributes are shared within lane / run; very few individual read-specific attributes. Are read names needed?
- How to ensure that essential metadata does get submitted?

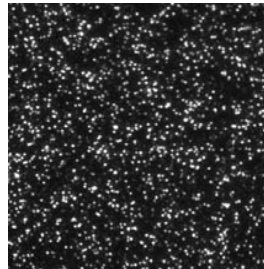
## Granularity – atomic units of retrieval



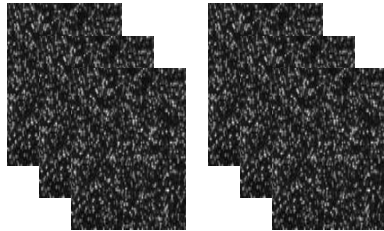
single read



a lane



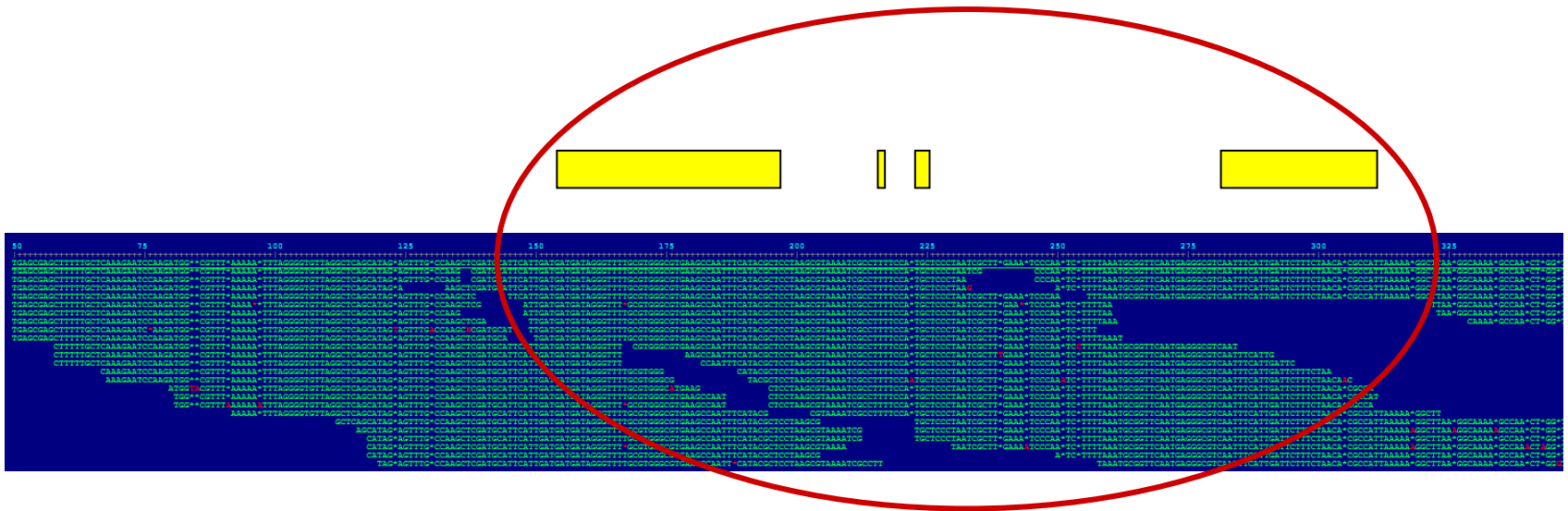
a run



multiple lanes/runs  
from an individual;  
project

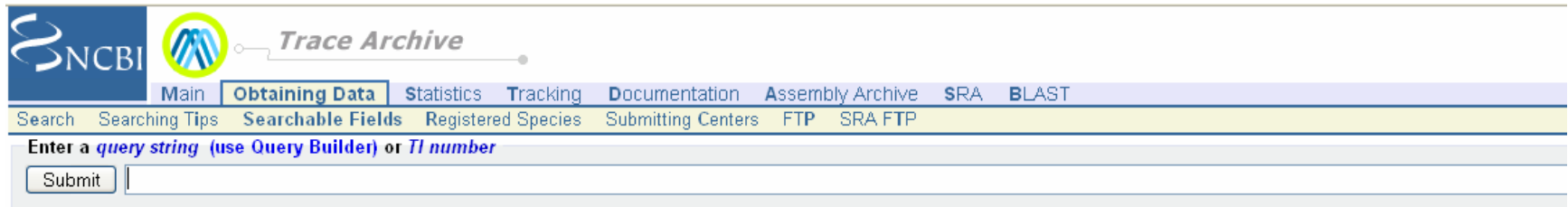
Would it make sense to break up the data into “sensible-sized” chunks (e.g. lane)?

# Context-driven retrieval?



Concessions to serve reads that align to a specific region (e.g. gene) potentially from a number of different runs/lanes?

# Data presentation – searching and browsing



NCBI Trace Archive

Main Obtaining Data Statistics Tracking Documentation Assembly Archive SRA BLAST

Search Searching Tips Searchable Fields Registered Species Submitting Centers FTP SRA FTP

Enter a *query string* (use *Query Builder*) or *T1 number*

Submit

## Searchable Fields



Name	Description
<a href="#">[see RFC] ACCESSION</a>	Genbank/EMBL/DDBJ accession number
<a href="#">[see RFC] AMPLIFICATION_FORWARD</a>	The forward amplification primer sequence
<a href="#">[see RFC] AMPLIFICATION_REVERSE</a>	The reverse amplification primer sequence
<a href="#">[see RFC] ANONYMIZED_ID</a>	Anonymous ID for an individual
<a href="#">[see RFC] BASES_20</a>	Number of quality scores which exceed 20
<a href="#">[see RFC] BASES_40</a>	Number of quality scores which exceed 40
<a href="#">[see RFC] BASES_60</a>	Number of quality scores which exceed 60
<a href="#">[see RFC] CENTER_NAME</a>	Name of the sequencing center
<a href="#">[see RFC] CENTER_PROJECT</a>	Center defined project name
<a href="#">[see RFC] CHEMISTRY</a>	Description of the chemistry used in the sequencing reaction
<a href="#">[see RFC] CHEMISTRY_TYPE</a>	Type of chemistry used in the sequencing reaction
<a href="#">[see RFC] CHROMOSOME</a>	Chromosome to which the trace is assigned
<a href="#">[see RFC] CLIP_QUALITY_LEFT</a>	Left clip of the read, in base pairs, based on quality analysis
<a href="#">[see RFC] CLIP_QUALITY_RIGHT</a>	Right clip of the read, in base pairs, based on quality analysis
<a href="#">[see RFC] CLIP_VECTOR_LEFT</a>	Left clip of the read, in base pairs, based on vector sequence
<a href="#">[see RFC] CLIP_VECTOR_RIGHT</a>	Right clip of the read, in base pairs, based on vector sequence
<a href="#">[see RFC] CLONE_ID</a>	The name of the clone from which the trace was derived
<a href="#">[see RFC] COLLECTION_DATE</a>	The full date, in "Mar 2 2006 12:00AM" format, on which an environmental sample was collected
<a href="#">[see RFC] CVECTOR_ACCESSION</a>	Repository ( GenBank/EMBL/DDBJ) accession identifier for the cloning vector

Current structure is centered on individual traces... but many attributes can be directly applied to batches (lanes, runs) of reads

How will data users find relevant datasets?

# Retrieval mechanisms

- Web-based



Trace Archive

[Main](#)
[Obtaining Data](#)
[Statistics](#)
[Tracking](#)
[Documents](#)

[Search](#)
[Searching Tips](#)
[Searchable Fields](#)
[Registered Species](#)
[Submitting](#)

Enter a *query string* (use [Query Builder](#)) or *TI number*

## Index of ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead

[Up to higher level directory](#)

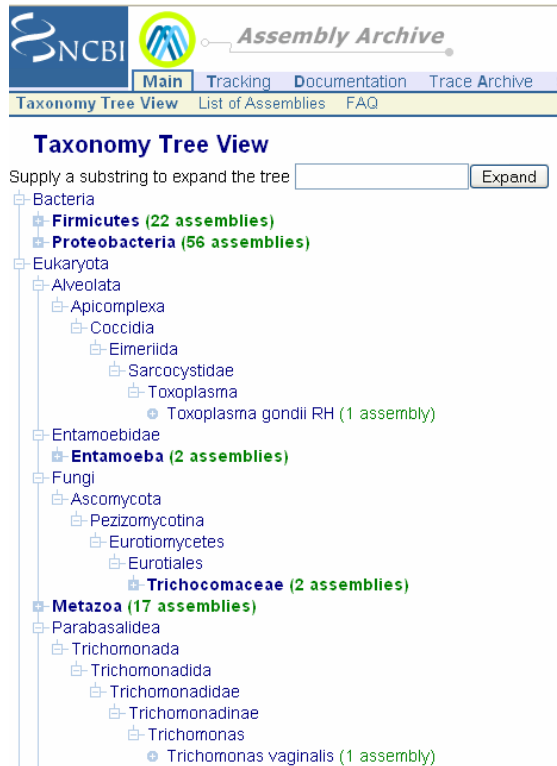
<a href="#">BCM</a>	7/11/2007 7:38:00 PM
<a href="#">CSHL</a>	7/11/2007 9:07:00 PM
<a href="#">JGI</a>	7/17/2007 3:25:00 PM
<a href="#">...</a>	...

- Programmatic (tied in with data views?)

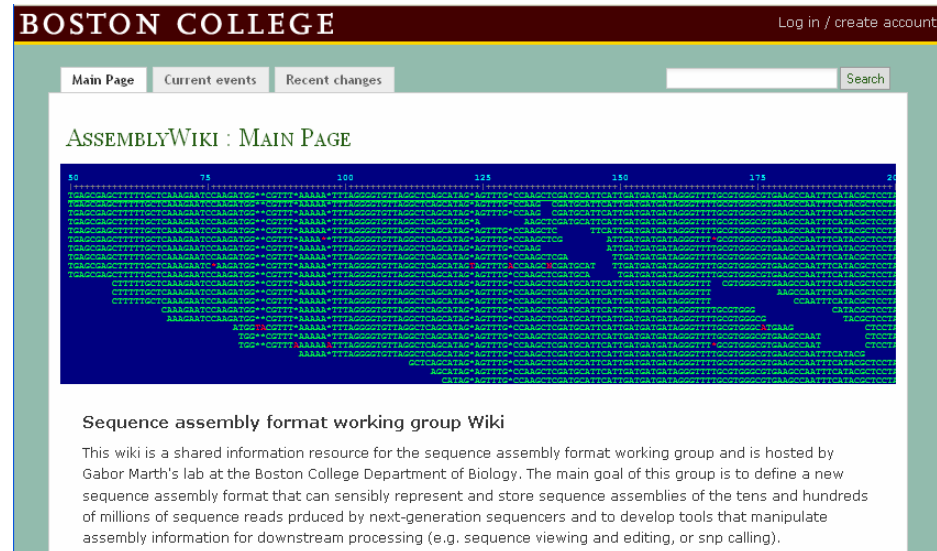
[illegible]



# Connection to assembly archive & data formats



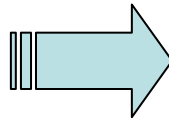
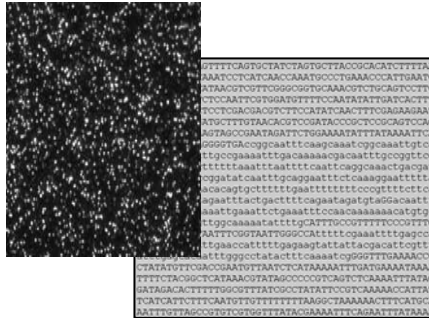
The screenshot shows the NCBI Assembly Archive website. The top navigation bar includes links for Main, Tracking, Documentation, and Trace Archive. Below this is a 'Taxonomy Tree View' section with a search input and an 'Expand' button. The tree is rooted at 'Bacteria' and branches into 'Firmicutes (22 assemblies)', 'Proteobacteria (56 assemblies)', 'Eukaryota', and 'Metazoa (17 assemblies)'. The 'Eukaryota' branch is expanded, showing 'Alveolata', 'Entamoebidae', 'Fungi', and 'Metazoa'. The 'Fungi' branch is further expanded, showing 'Ascomycota', 'Eurotiomycetes', and 'Trichocomaceae (2 assemblies)'. The 'Metazoa' branch is also expanded, showing 'Parabasalidea', 'Trichomonada', and 'Trichomonadidae'. The 'Trichomonadidae' branch is further expanded, showing 'Trichomonadinae' and 'Trichomonas'. The 'Trichomonas' branch is further expanded, showing 'Trichomonas vaginalis (1 assembly)'.



The screenshot shows the BOSTON COLLEGE ASSEMBLYWIKI : MAIN PAGE. The page has a dark red header with the text 'BOSTON COLLEGE' and a 'Log in / create account' link. Below the header is a navigation bar with links for 'Main Page', 'Current events', and 'Recent changes'. The main content area is titled 'ASSEMBLYWIKI : MAIN PAGE' and contains a large block of sequence data in FASTQ format. The sequence data is displayed in a monospaced font with line numbers 10, 75, 100, 125, 150, and 175. The sequence data is a mix of uppercase and lowercase letters, representing nucleotide bases. Below the sequence data is a section titled 'Sequence assembly format working group Wiki' with a paragraph of text: 'This wiki is a shared information resource for the sequence assembly format working group and is hosted by Gabor Marth's lab at the Boston College Department of Biology. The main goal of this group is to define a new sequence assembly format that can sensibly represent and store sequence assemblies of the tens and hundreds of millions of sequence reads produced by next-generation sequencers and to develop tools that manipulate assembly information for downstream processing (e.g. sequence viewing and editing, or snp calling)'.

- How to make the connection between reads in the read archive and the assembly archive? ➔ **Back to unique read IDs**

# Clientele – who will this archive cater to?



- Application support for short-read data manipulation (data access libraries)?