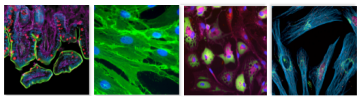


Can Python Save Next Generation Sequencing?

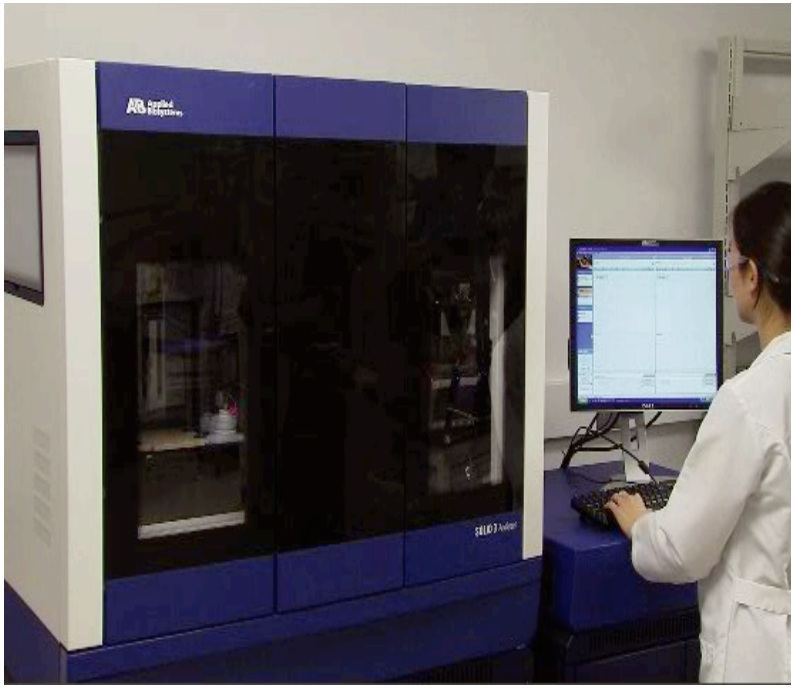
Chris Mueller
Life Technologies
June 30, 2010
SciPy 2010 Austin, TX

life technologies™

Human Genome Project

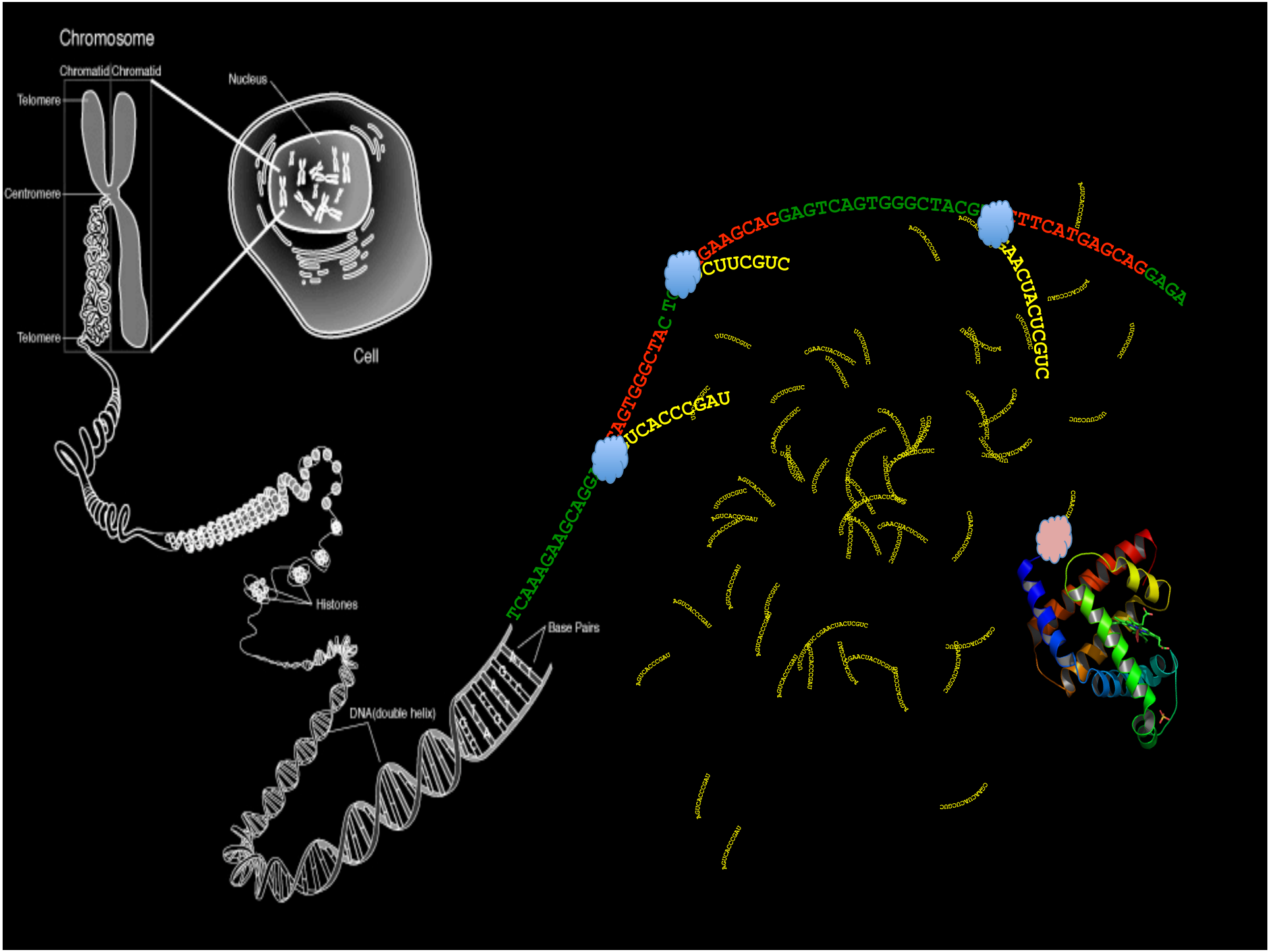
10 Years
Thousands of Sequencers
$3,000,000,000
21 Billion Base Pairs (Gbp)

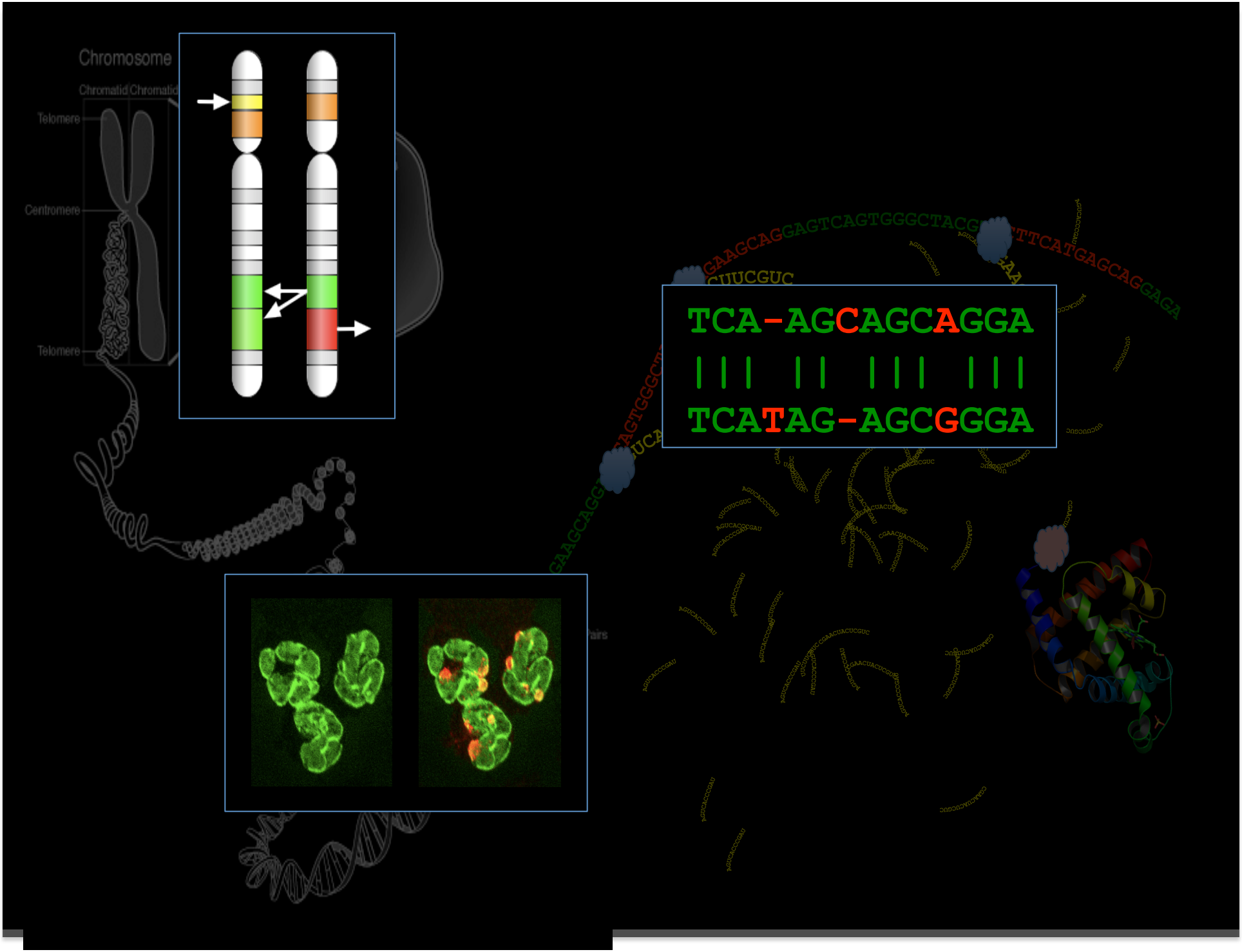
Modern Sequencing



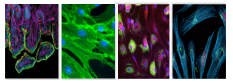
2 weeks
One Sequencer
$6,000
100-200 Gbp

life technologies™

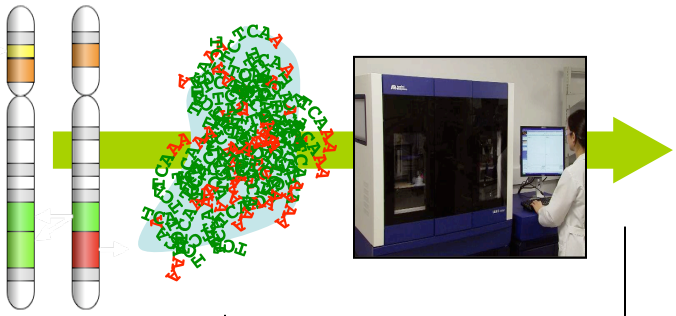

Chromosome
Chromatid
Chromatid
Telomere
Centromere
Telomere
Nucleus
Cell
Histones
DNA(double helix)
Base Pairs
TCAAAGAAGCAGG
GAAGCAGGAGTCAGTGGGCTACG
TTCATGAGCAGGAGA
CUUCGUC
UCACCCGAU
AACUACUCGUC


Chromosome
Chromatid Chromatid
Telomere
Centromere
Telomere
TCA-AGCAGCAGGA
TCATAG-AGCGGGA

Next Generation Sequencing (NGS)



Next Generation Sequencing (NGS) instruments such as Life Technologies SOLiD™ sequence genomic material in millions of small pieces, enabling a high-level of throughput and sequencing depth.

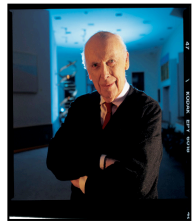


Sample prep breaks the DNA or RNA into short segments that are attached 500 million to 1 billion beads.

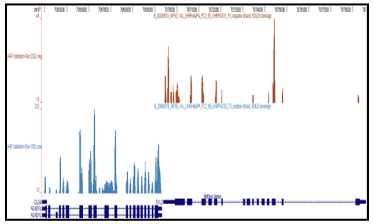
10 GB – 2 TB of raw data is transferred from the instrument to the analysis cluster.

```
.
.
>187_29_706_F3
T233020103031311231230222031111232002101001220011 02
>187_29_800_F3
T311200122132220022221301211211221120322203231212 02
>187_29_824_F3
T222111300230201332313233023103031311231220221121 1
>187_29_829_F3
T233020100031301231230222031111201221232021321230 1212
>187_29_858_F3
T233020103031311231230222031111232221231221232121 212
>187_29_885_F3
T233020103031311231230222031111212200132121220212 22
.
.
```

The sequence from each bead is reported in a data file. Files can be over 100 GB.



The bead sequences are either assembled into a new genome or “mapped” to a reference genome.



The mapped reads from RNA samples can be further analyzed to determine which genes are active in the sample.

life technologies™

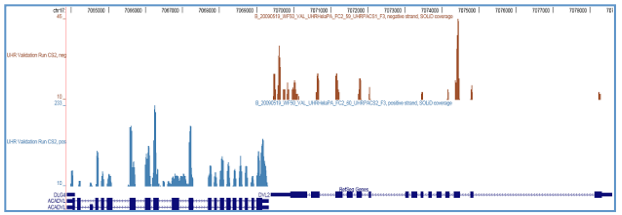
The Goal: Medical Genomics

Acquire a Reference

Sequence the diseased genome

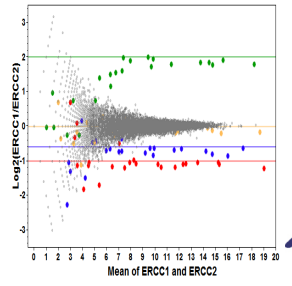
```
GATCAACTTG AGGCCAGCCT GACCAACGTG GCAAAACTCC ATCTCTACTA
AATTAGCTGA GCGTGGTGGC ACGCACCTGT CATCCCAGTT ACTCAGGAGG
AGAATGGCTT GAGCCTGGGA GACAGAGGTT GCAGTGAGCT GAGATTGCAT
TAGCCTGGGT GACAGAGTGA AATGGAGGGA GGAAAAAAAA AAAAAAGGAA
AGGGAGCCAG CCTAGGATGG GGAAGGCTCA CCAGAAGTGG ATGCAAAGAG
GAGCTCATTC TATTTGCCTA GGAAAGAAAA ACGTCCAGAA ACCTGGCCTT
GCCGAGGCCC TCCAGGAAAG CCAGGCAGAC CCTGCTCCTG CTCTGACCCC
```

Sequence the expressed RNA



Annotate the diseased genome

Compare against the healthy sample

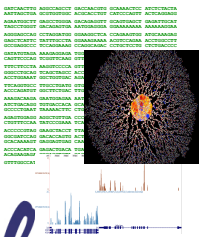


life technologies™

Personal Genomics

Acquire a Reference	10 GB	2-3 HD Movies
Sequence the diseased genome	100 Gbp 36 Hours/4 servers	6% of Google's Index
Sequence the expressed RNA	100 Gbp 15 Billion Search Ops	30x Google's Daily Search Traffic
Annotate the diseased genome	Millions of Features	50-500 GB Retained Data
Compare against the healthy sample	Human Driven Analysis and Interpretation	

vs.



life technologies™

Population Genomics

Acquire a Reference	1 Petabyte of Bandwidth	1 Billion YouTube Videos
Sequence the diseased genome	4 Million Servers	1.2 million servers were sold in Q4 2009
Sequence the expressed RNA	72 Million Hours	3 Years on UT's Ranger Supercomputer
Annotate the diseased genome	1 Exabyte/Day	The Large Hadron Collider generates 300 EB/year
Compare against the healthy sample	27,000 Radiologists in the US	

life technologies™

So…

Can Python save next generation sequencing?

Probably not on its own…

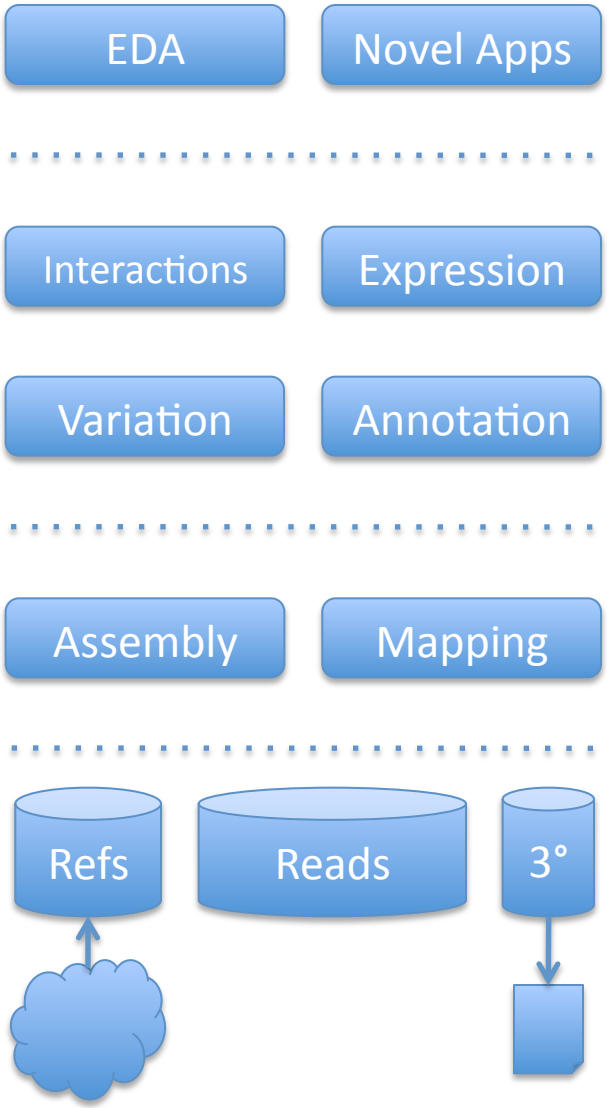
Better questions:

How can Python help next generation sequencing?

How can next generation sequencing help Python?

life technologies™

NGS Data and Workflow Components



Exploratory Analysis

Undirected analysis and unforeseen sequencing applications

Standard Scientific Workflows

Understand the structure and function of genomic elements

Fundamental Algorithms

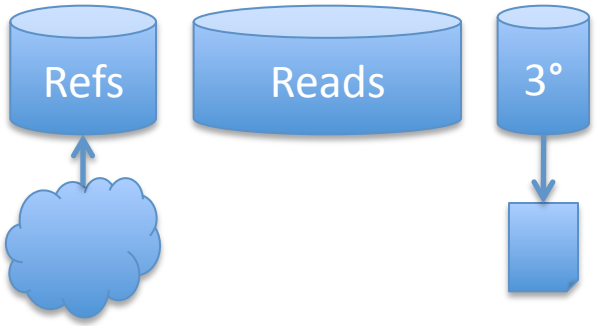
Transform the raw data into scientifically relevant forms

References and Data

Reference genomes, domain-specific data sets, raw data, analysis results

life technologies™

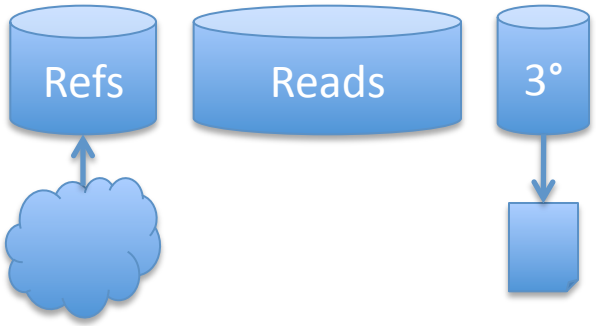
Algorithms and Scale

Data and Workflows	Algorithms	Considerations
EDA, Novel Apps	Information visualization, data-mining	Interactivity is essential.
Interactions, Expression, Variation, Annotation	Clustering, statistical models, network analysis	Data is smaller, but analysis may require round trips back to reads.
Assembly, Mapping	Graphs, Index schemes, dynamic programming	Generally I/O bound with opportunities for parallelism.
	Standard data formats are text-based	Can easily span 100s of TBs for a small lab. References are in constant flux.

life technologies™

Software and Hardware

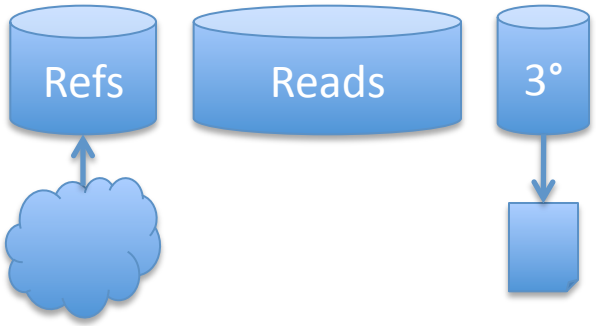
Data and Workflows	Software	Hardware
EDA, Novel Apps	Genome browsers, Matlab, Stats tools, R, etc	Workstations, laptops with fast access to storage
Interactions, Expression, Variation, Annotation	Scripting languages, data analysis libraries	Single nodes with fast access to storage
Assembly, Mapping	Graph libraries, pipeline management software, job schedulers	Clusters with high-memory nodes, fast access to storage.
Refs, Reads, 3°	Databases, ORM tools, **Flat files**	Distributed file systems, Fast SANs



life technologies™

Python and NGS

Data and Workflows	Current	Potential
EDA, Novel Apps	Galaxy, user-developed tools	GUIs, Disco
Interactions, Expression, Variation, Annotation	User tools, utilities shipped with assemblers and mappers, NGS libraries (HTSeq), NumPy, SciPy	More libraries!
Assembly, Mapping	Rapid prototyping, pipeline management, utilities	Multi-processing, queuing to build out pipeline manager
Refs, Reads, 3°	Python parsers for various formats, BioPython, SAMTools Pygr for reference and annotation access	Disco or Hadoop for distributed read management



life technologies™

Example: Chromosome 20 Expression Analysis

A portion of the Chromosome 20 expression map created using reads from SOLiD™ at Life Technolgies' Austin site. The full map spans over 260 feet when printed sequentially. Chromosome 20 represents 2.1% of the human genome.

Visualization — Custom Python tool rendering to PDF using ReportLab

Expression — LifeTech WT pipeline for mapping and expression analysis

Mapping — LifeTech WT pipeline for mapping and expression analysis

Annotation — Custom Python scripts for result aggregation

Reads — 14 Billion reads from three sample types. 1400+ files comprising 4TB of sequence data

Refs — Human Genome 36.3
Human RefSeq 39

life technologies™

Cultural Challenges

As recently as five years ago, computing in the Life Sciences consisted of pencils, lab notebooks, and the occasional Excel spreadsheet. Needless to day, NGS caught the community completely off guard.

- Goals
  - Most users care about the science and just want the technology to work
- Skills
  - Few biologists and bioinformaticians are also experts in HPC and tera-scale data mining
- Expectations
  - Excel provides answers instantly. NGS analysis should, too.
- Cost
  - I just spent $500k on an instrument and I need to spend how much on a computer and developer???

life technologies™

Final Thoughts

- NGS analysis is in its infancy
  - Fundamental methods are still being developed
  - The scientific community is still coming to grips with its potential and challenges
  - There won't be any silver bullets in the near future
- NGS analysis is complicated by the scale of data
  - Traditional supercomputing doesn't help much
- Development paradigms that simplify large data processing will succeed in this space
- Many Python projects show promise, but there's still work to be done!

life technologies™

Thank You!

Life Technologies Austin

Bioinformatics
Jeff Schageman
Joel Brockman
Penn Whitley
Dan Williams

Transcriptomics
Sheila Heater
Kelli Bramlett
Diane Ilsley

The Powers that Be
Bob Setterquist
Tim Sendera

Life Technologies Global

Software and HPC
Lee Jones
Patrick LeGresley
Somalee Datta
Asim Siddiqui
Aaron Kitzmiller
Keith Moulton
Mike Lyons
Ying Zu

IT
Michael Moore
Antoine Uzzeni
David Morgan
Joanna Curlee

Sounding Boards

Peter Wang
Glen Otero
Travis Oliphant

life technologies™

What People are Using

- Compiled Languages:
  - Equally distributed between C, C++, and Java
- Interpreted Languages:
  - Almost everyone uses Perl, with about half the respondants using R or Python
  - But, the for daily usage, about 31% used Python and 47% used Perl
- For Libraries:
  - 75% used BioPerl on a regular basis. SciPy is the next most used at 33%. BioPythyon was low at 16%
- For Statistics:
  - R (70%), Excel (53%), and Perl (35%) were the most common
- For Visualization:
  - Excel was the most common with R and GNUPlot also getting multiple responses

life technologies™

Compute Environment

- Peta-scale storage environment
- Multi-core processors and multiple nodes
  - though not as many as you think
- 64+ GB of RAM helps
- Heterogeneous processors
  - SIMD/GPUs
- Multiple languages
  - C, C++, and Java are common for 'fast' code
  - Many tools ship with Perl or Python utilities
- The cloud will matter at some point

life technologies™

Ideal Stack

- Must be useable by a wide range of developers and scientists
- Standard data formats
- Reference Management
- Optimized kernel operations
- Optimized record-based operations
- Cluster aware pipeline management
- Horizontally scalable record storage
- Database integration for LIMs
- Interactive visualization

life technologies™