SEQanswers.com:

the next generation sequencing community

Eric C. Olivares, Ph.D.
eric@seqanswers.com
Twitter: @SEQanswers
May 25, 2011

Bay Area Bioinformatics Meetup

Overview

- Introduction to me
- Introduction to SEQanswers.com

- General library preparation
- Sequencing technologies

- Future of the site
- Questions

Who am I?

- 1998 UC Davis, BS, Genetics
- 2003 Stanford, PhD, Genetics
  - "Phage integrases for gene therapy: from concept to applications"
- 2003 Founding scientist of startup

- 2005 Invitrogen R&D
  - PureLink™ Genomic DNA
  - RiboMinus™ system
- 2007 Founded SEQanswers.com

- 2008 Pacific Biosciences R&D
  - Sample prep

A note about perceived conflicts…

- I work at and own stock in Pacific Biosciences
  - Sponsors of this session
  - Didn't edit my talk
- I also run SEQanswers
  - Advertisers…thanks!
- Call bias if you see it, I'm human

Why SEQanswers.com?

- 2007, Nothing Existed
- Community is key
  - News
  - Opinions
  - Technology discussions
  - Protocol/experience sharing
  - Software tools
- Local is good, worldwide is better
- Hobbies demand forums!





What is SEQanswers.com?

1. Forum style information resource

2. Wiki organizing more defined information
   - Software tools
   - Sequencing service providers

Forum style community information resource

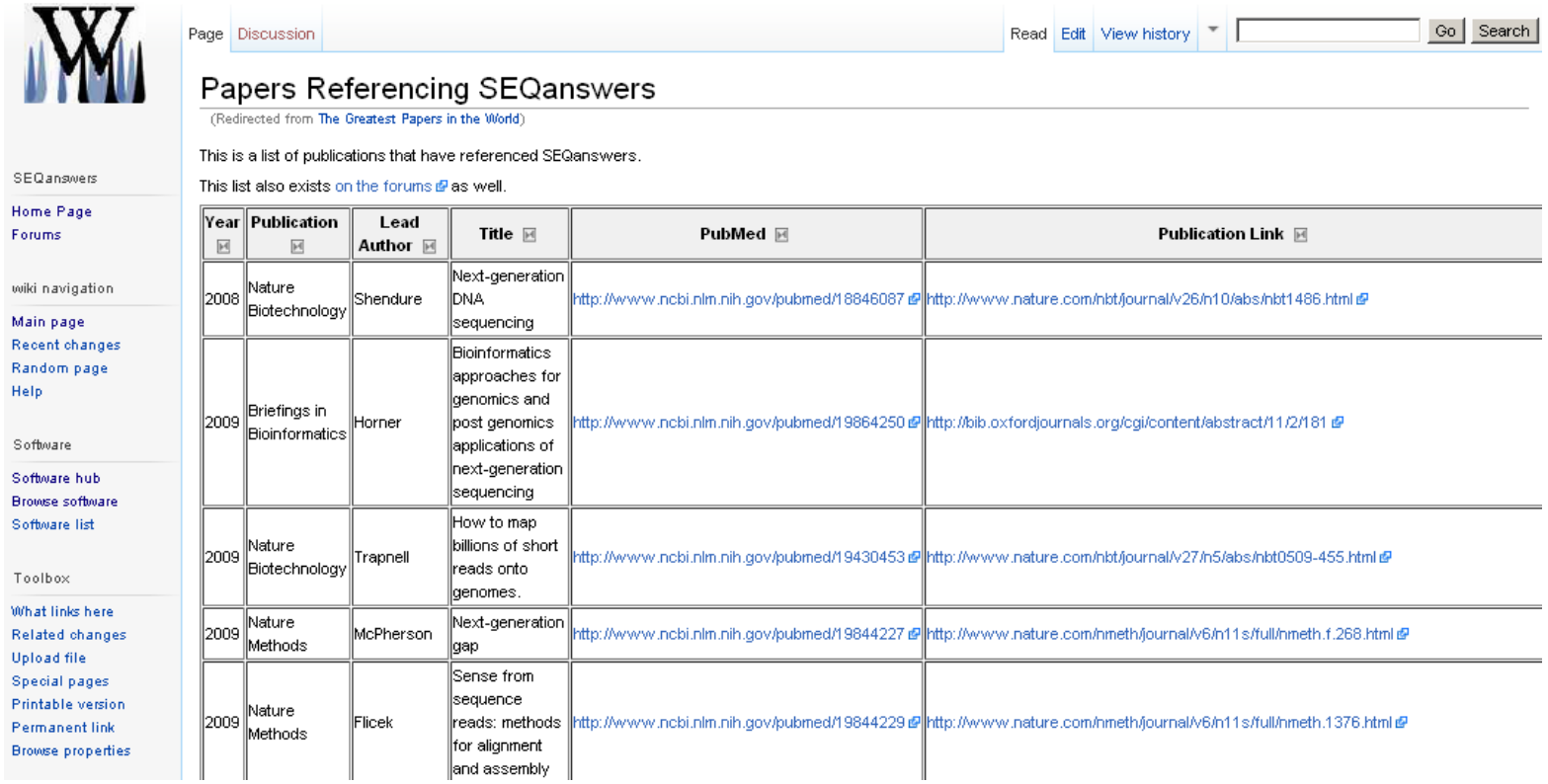
- 30 subforums
- *Bioinformatics*
- Supplier specific forums (ILMN, 454, etc)
- Application specific forums (Sample Prep, RNA-seq, etc)
- Literature Watch
- Events/Conferences
- Jobs forums

The power of community…

- Software troubleshooting at its best…
  - "SEQanswers boasts a number of authors of major second generation sequencing analysis packages, including Bowtie, Tophat, BFAST and Samtools, as regular contributors." –Keith Robison, OmicsOmics.blogspot.com
    http://omicsomics.blogspot.com/2011/05/forums-open-beats-closed-hands-down.html

    - Bowtie = Ben Langmead (u=1129) 194 posts
    - Tophat = Cole Trapnell (u=1564) 183 posts
    - BFAST = Nils Homer (u=1586) 895 posts (moderator!)
    - Samtools = Heng Li (u=129) 449 posts
    - Many, many others…

14 appearances in the literature…



Page | Discussion | Read | Edit | View history | Go | Search

Papers Referencing SEQanswers

(Redirected from The Greatest Papers in the World)

This is a list of publications that have referenced SEQanswers.

This list also exists on the forums as well.

SEQanswers
Home Page
Forums

wiki navigation
Main page
Recent changes
Random page
Help

Software
Software hub
Browse software
Software list

Toolbox
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Browse properties

Year	Publication	Lead Author	Title	PubMed	Publication Link
2008	Nature Biotechnology	Shendure	Next-generation DNA sequencing	http://www.ncbi.nlm.nih.gov/pubmed/18846087	http://www.nature.com/nbt/journal/v26/n10/abs/nbt1486.html
2009	Briefings in Bioinformatics	Horner	Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing	http://www.ncbi.nlm.nih.gov/pubmed/19864250	http://bib.oxfordjournals.org/cgi/content/abstract/11/2/181
2009	Nature Biotechnology	Trapnell	How to map billions of short reads onto genomes.	http://www.ncbi.nlm.nih.gov/pubmed/19430453	http://www.nature.com/nbt/journal/v27/n5/abs/nbt0509-455.html
2009	Nature Methods	McPherson	Next-generation gap	http://www.ncbi.nlm.nih.gov/pubmed/19844227	http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.f.268.html
2009	Nature Methods	Flicek	Sense from sequence reads: methods for alignment and assembly	http://www.ncbi.nlm.nih.gov/pubmed/19844229	http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1376.html

SEQwiki

- Organizing definable information
  - Software tools
  - Sequencing service providers

- Community edited

Immense gratitude owed to Dan Bolser, post doc @ University of Dundee, Scotlar

Forums are terrible at some things

Sticky: Software packages for next gen sequence analysis (1 2 3) sci_guy	12-25-2009 06:45 PM by ECO	236	162,503		

01-23-2008, 11:19 PM #1

sci_guy
Member

Join Date: Jan 2008
Location: Sydney
Posts: 76

Software packages for next gen sequence analysis

28 Dec 2009: This thread has been closed. Please see our wiki software portal for information about each of these packages.

A reasonably thorough table of next-gen-seq software available in the commercial and public domain

Integrated solutions

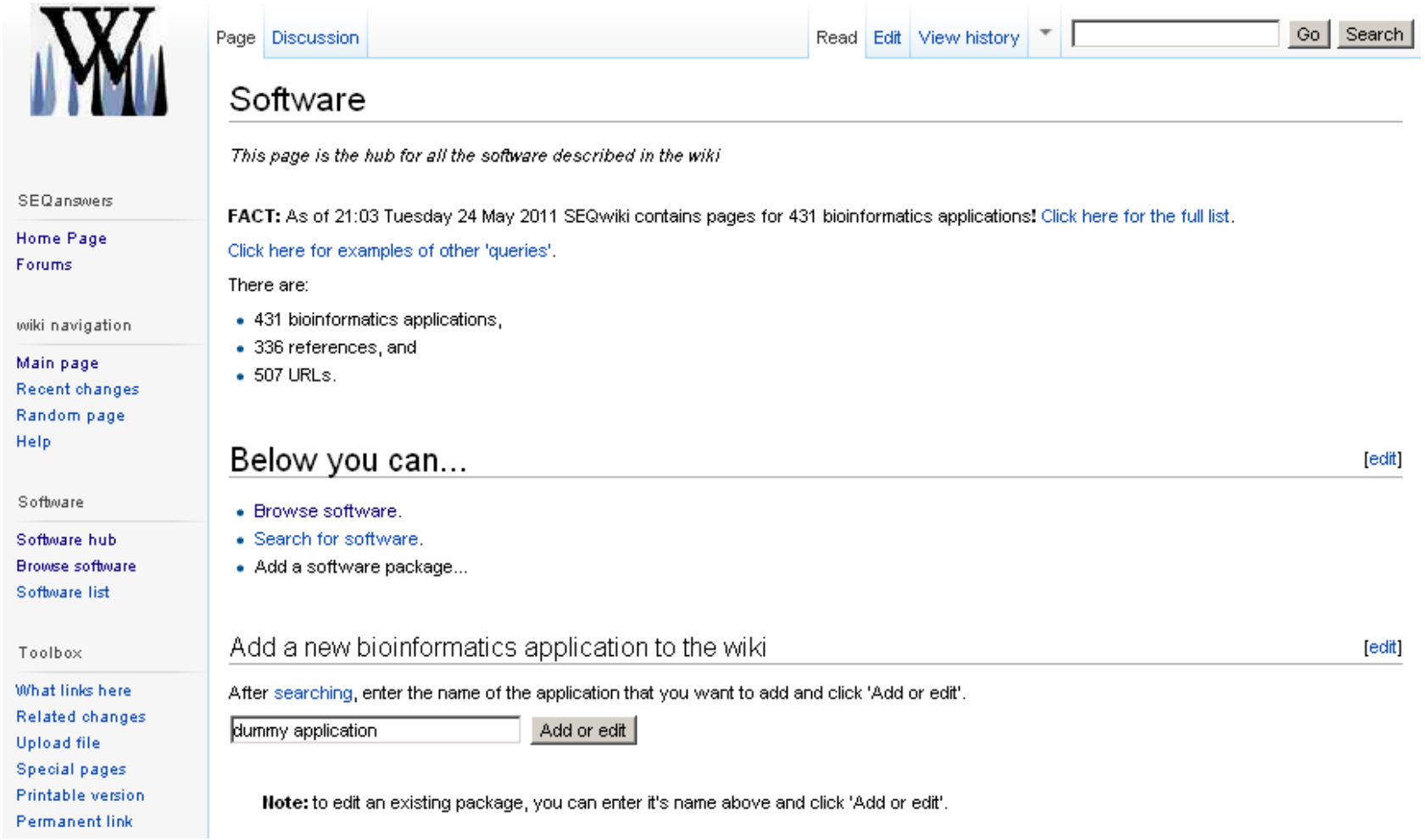
* CLCbio Genomics Workbench - *de novo* and reference assembly of Sanger, Roche FLX, Illumina, Helicos, and SOLiD data. Commercial next-gen-seq software that extends the CLCbio Main Workbench software. Includes SNP detection, CHiP-seq, browser and other features. Commercial. Windows, Mac OS X and Linux.
* Galaxy - Galaxy = interactive and reproducible genomics. A job webportal.
* Genomatix - Integrated Solutions for Next Generation Sequencing data analysis.
* JMP Genomics - Next gen visualization and statistics tool from SAS. They are working with NCGR to refine this tool and produce others.
* NextGENe - *de novo* and reference assembly of Illumina, SOLiD and Roche FLX data. Uses a novel Condensation Assembly Tool approach where reads are joined via "anchors" into mini-contigs before assembly. Includes SNP detection, CHiP-seq, browser and other features. Commercial. Win or MacOS.
* SeqMan Genome Analyser - Software for Next Generation sequence assembly of Illumina, Roche FLX and Sanger data integrating with Lasergene Sequence Analysis software for additional analysis and visualization capabilities. Can use a hybrid templated/de novo approach. Commercial. Win or Mac OS X.
* SHORE - SHORE, for Short Read, is a mapping and analysis pipeline for short DNA sequences produced on a Illumina Genome Analyzer. A suite created by the 1001 Genomes project. Source for POSIX.
* SlimSearch - Fledgling commercial product.

Align/Assemble to a reference

* BFAST - Blat-like Fast Accurate Search Tool. Written by Nils Homer, Stanley F. Nelson and Barry Merriman at UCLA.
* Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a typical workstation with 2 gigabytes of memory. Uses a Burrows-Wheeler-Transformed (BWT) index. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell. Linux, Windows, and Mac OS X.
* BWA - Heng Lee's BWT Alignment program - a progression from Maq. BWA is a fast light-weighted tool that aligns short sequences to a sequence database, such as the human reference genome. By default, BWA finds an alignment within edit distance 2 to the query sequence. C++ source.
* ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony J. Cox for the Solexa 1G machine.
* Exonerate - Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney from EMBL. C for POSIX.

SEQwiki Software Hub

http://seqanswers.com/wiki/Software



Yup, still terrible…

12-16-2008, 07:59 PM #1

ECO
--Site Admin--

Join Date: Oct 2007
Location: SF Bay Area, CA, USA
Posts: 824

List of Next Gen Sequencing Service Providers

Let's start one! I've got some in a list but figured we could all do it quicker, so I'm opening this one up to everyone!

29 Dec 08: Some great work by everyone, I've started to compile it below. Please let me know if you have comments, corrections, or additions to anything. Everything is listed in region-company alphabetical order to remain fair.

21 May 09: Many additions thanks to everyone who posted here, and to mjleaks for compiling them for me!

28 Oct 09: Many more additions, added states and alphabetized the US. Sincerest apologies for even *insinuating* that NZ was part of the Australian continent. Oceania FTW.

9 May 10: Updated. Data ported to Wiki!

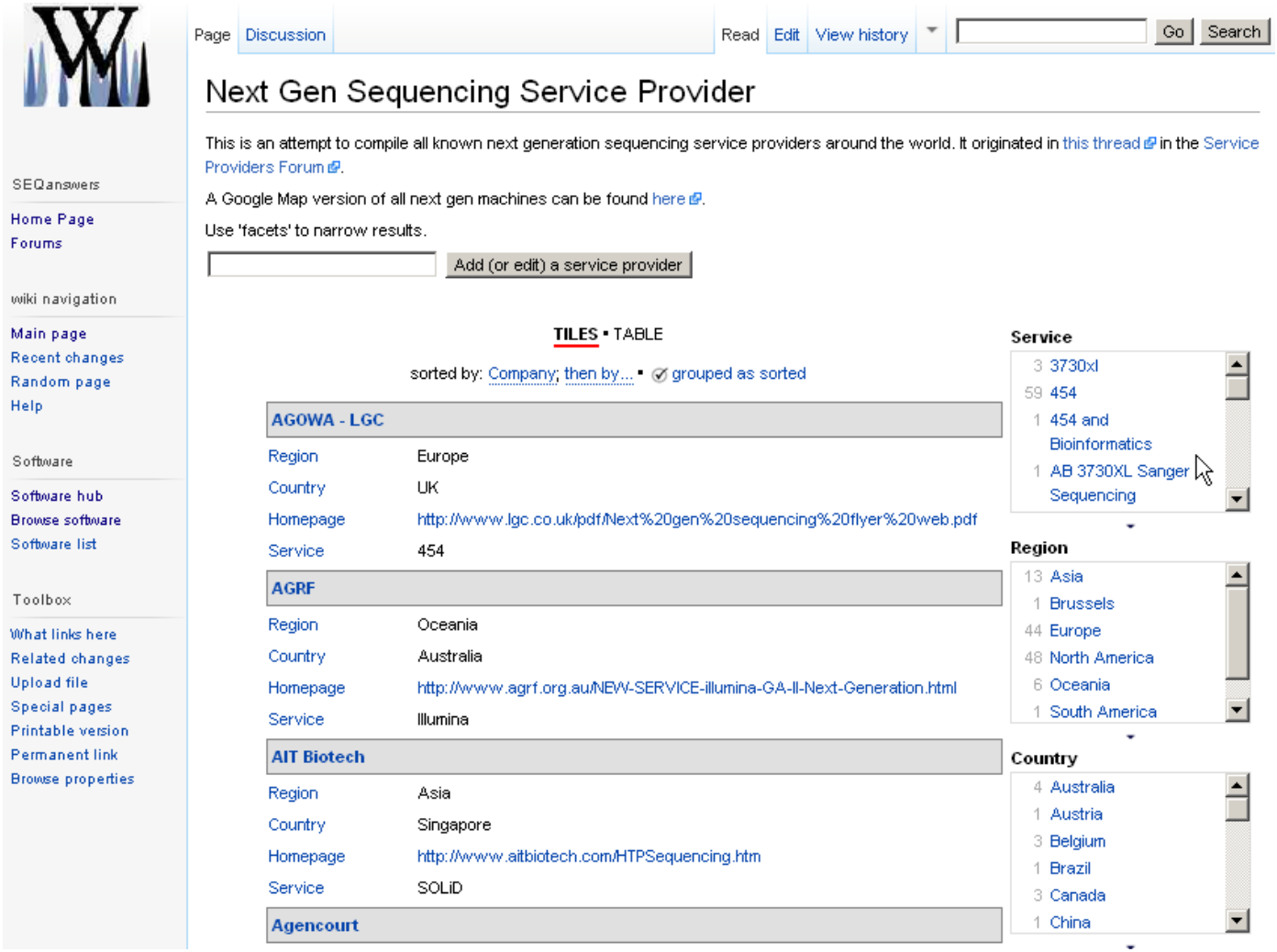
See and edit the same data in the wiki:

Next_Gen_Sequencing_Service_Providers

Region	Country	Company	Homepage	Instruments
Asia	China	Beijing Genomics Institute	http://www.genomics.org.cn/en/index.php	454 / Illumina (*only* 30 GAs) / SOLiD
	India	Genotypic	http://genotypic.co.in/newgen/	454 / Helicos / Illumina / SOLiD
	India	Geschickten Solutions	http://www.geschickten.com/	Informatics
	India	Ocimum Biosolutions	http://www.ocimumbio.com/web/	454 / Illumina / Informatics / SOLiD
	India	Xcelris Labs	http://www.xcelrislabs.com/	454 / SOLiD
	Japan	Riken Genesis	http://www.rikengenesis.jp/en/index.html	??
	Korea	Macrogen	http://www.macrogen.com/	454 / Illumina / SOLiD
	Singapore	AIT Biotech	http://www.aitbiotech.com/HTPSequencing.htm	SOLiD
	Taiwan	MissionBio	http://www.missionbio.com.tw/	SOLiD
Oceania	Australia	AGRF	http://www.agrf.org.au/NEW-SERVICE-i...eneration.html	Illumina
	Australia	Geneworks	http://www.geneworks.com.au/Content.aspx?p=84	Illumina
	Australia	Micromon (Department of Microbiology, Monash University, Victoria)	http://dna.med.monash.edu.au/	Illumina / Informatics
	New Zealand (definitely not Australia)	Allan Wilson Centre Genome Service at Massey University	http://www.allanwilsoncentre.ac.nz/AWCGSintro.htm	Illumina
	New Zealand (definitely not Australia)	University of Otago	http://sequence.otago.ac.nz/index.html	454 / SOLiD

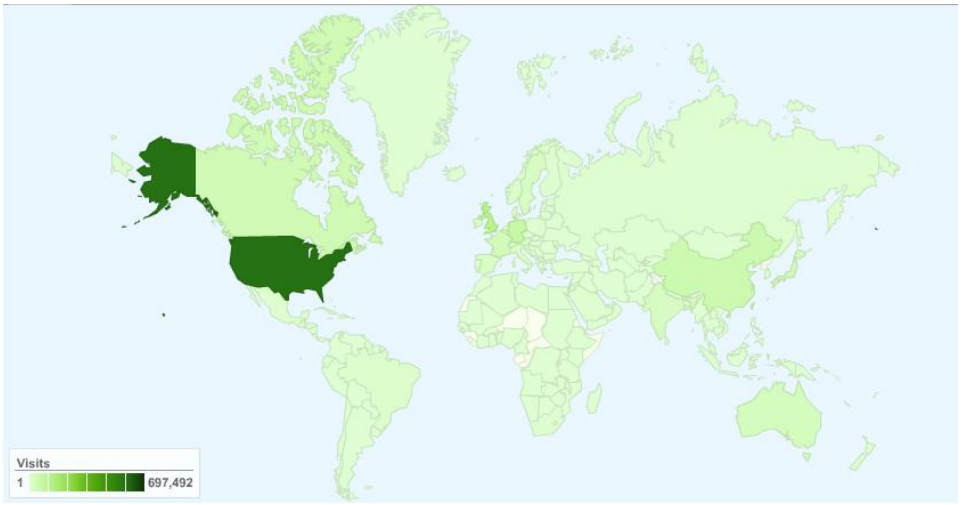
SEQwiki NGS Service Providers

http://seqanswers.com/wiki/Next_Gen_Sequencing_Service_Provider

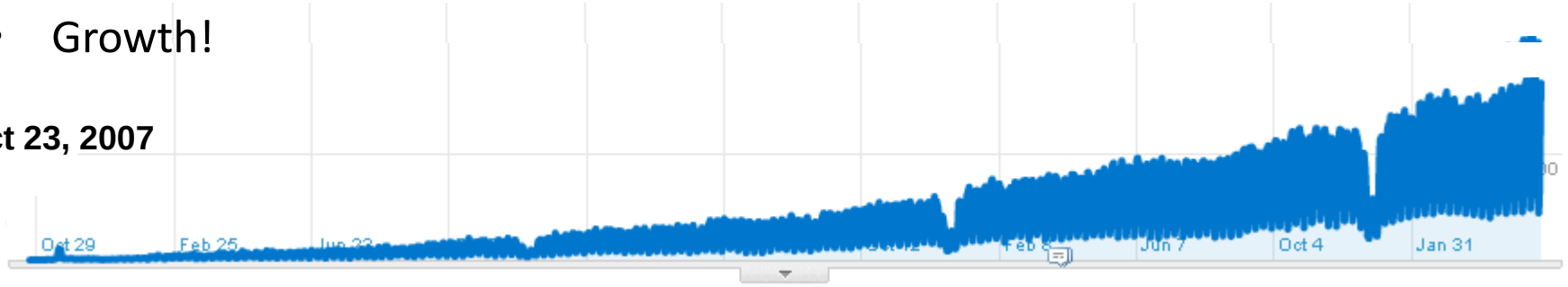


Traffic

- 1000s of unique worldwide visitors per day
- Hundreds of thousands of pageviews per month
- As of 5/24:
  - Threads: 9,425
  - Posts: 39,963
  - Members: 17,258
- Growth!



Oct 23, 2007



Library Prep Basics

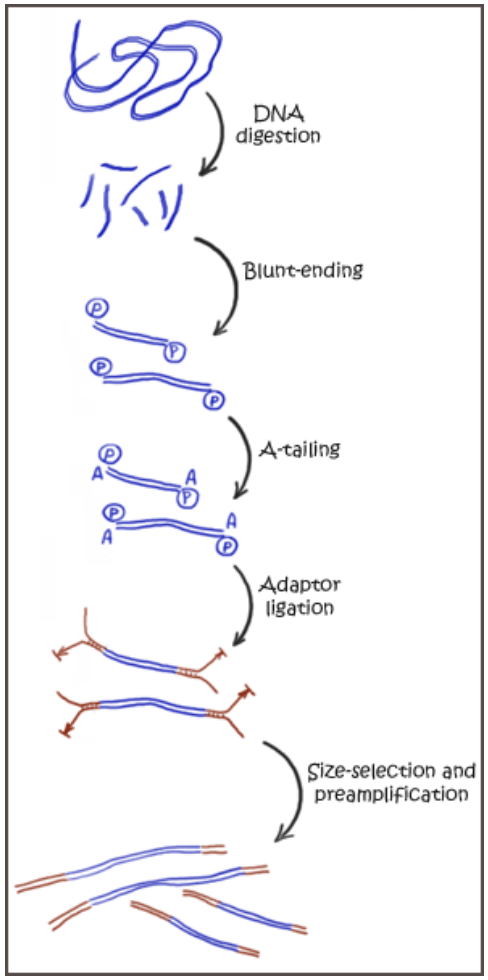
Library Prep Basics

- Library:
  - Representative pool of DNA fragments from sample of interest

- Ideal library:
  - Tight size distribution of fragments
  - Bounded by a known anchor for priming
  - Bias-free
    - Amplification
    - Shearing
    - Molecular biology (heat, denaturants, etc)

Generic Fragment Library Workflow

- Fragment DNA
  - Various physical or enzymatic methods
- Repair Ends
  - Fix staggered ends and add 5’ phosphates
- Ligate Adapters
  - Add small universal oligonucleotides
- Purify/Size Select
  - Remove non- or partially-ligated products
- Amplify
  - Clustering (Illumina)
  - Emulsion (454, SOLiD, Ion Torrent)



Illumina example from:
http://seq.molbiol.ru/

Library considerations for Bioinformaticians

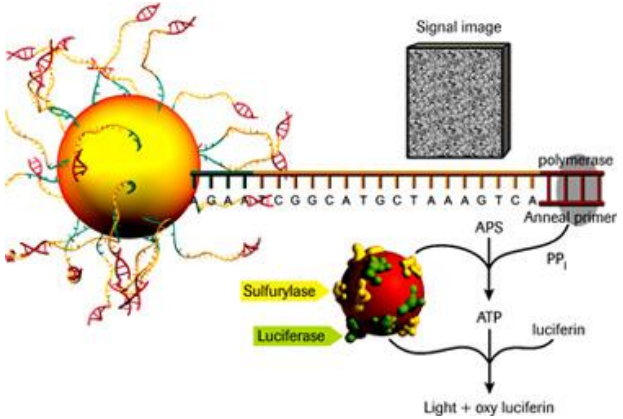
Problem	Symptom
GC bias	Uneven coverage, poor variant calls
Short readlengths	Mapping (repeats, structural variants, etc)
PCR duplicates	Artefactual variant calls
Chimeric inserts	Biologically irrelevant mappings
Low yield	Signal to noise, poor basecall quality
Wet-lab to biofx handoff	Incorrect data (Barcode assignment)
Contamination / Wrong Reference	Unaligned data

Sequencing Technologies

What's unique?

454: Pyrosequencing

- Pyrosequencing:
  - Release of pyrophosphate
  - Detection by enzymatic cascade which generates light
- Template beads held in PicoTiterPlate
- Chemistry creates homopolymer predominant error model
- Read lengths of 400 – 700bp

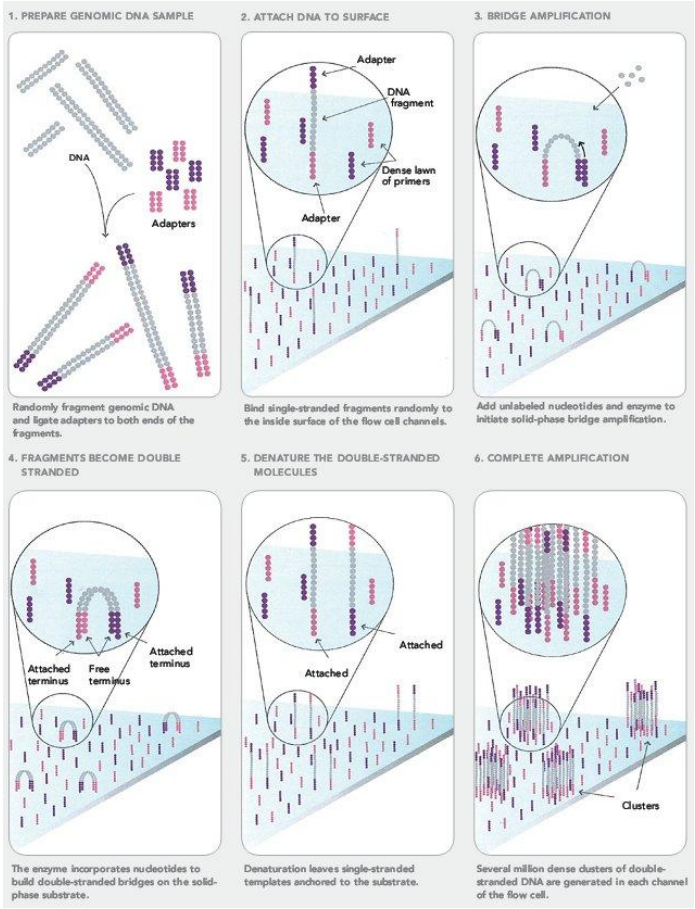


Illumina: Seq by Synthesis

- Sequencing by synthesis
  - Incorporate single labeled base
  - Image
  - Deprotect
  - Repeat

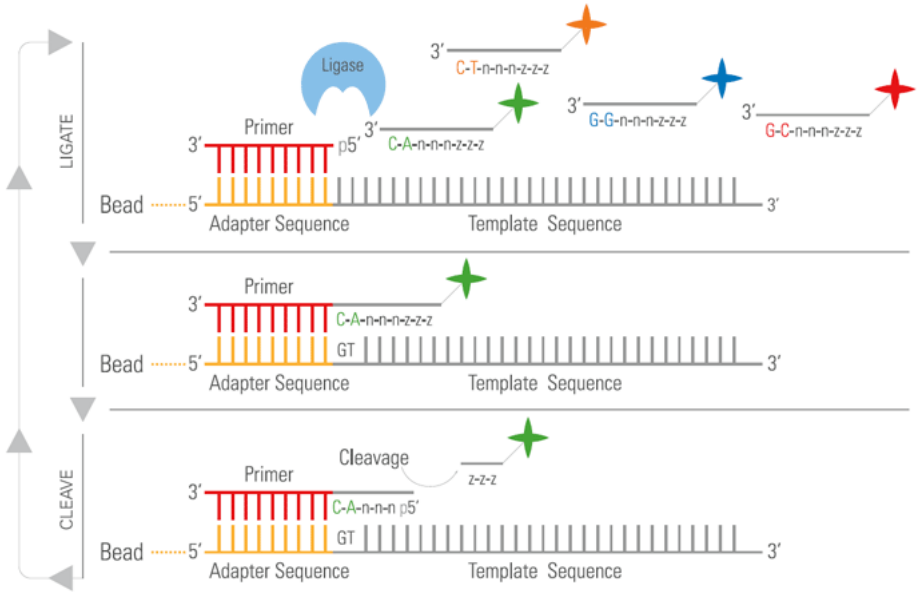
- Solid-phase immobilized clusters ('polonies') of templates on slide

- Readlengths of up to 150bp



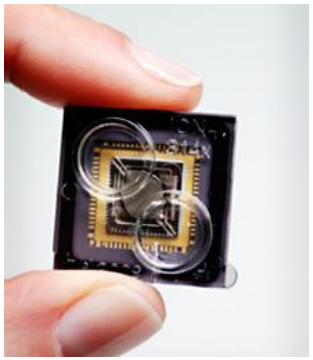
SOLiD: Seq by Ligation

- Sequential ligation of labeled probes
- Beads immobilized on glass slide
- Unique read property “colorspace”
- Interesting algorithmic challenges and market impacts



Ion Torrent PGM

- Founded by Jonathan Rothberg (454!)
- Direct electrical detection of protons released upon incorporation
- Beads immobilized in disposable parallelized microchip pH meter
- Very inexpensive
- Fast run times
- Readlengths <100bp

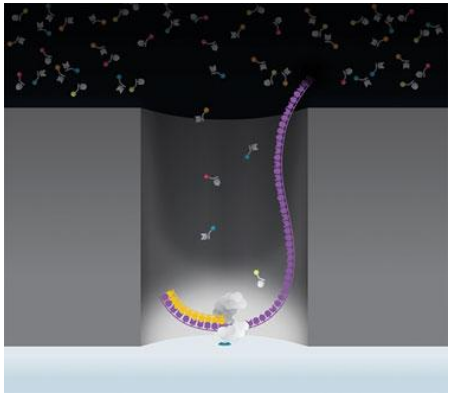


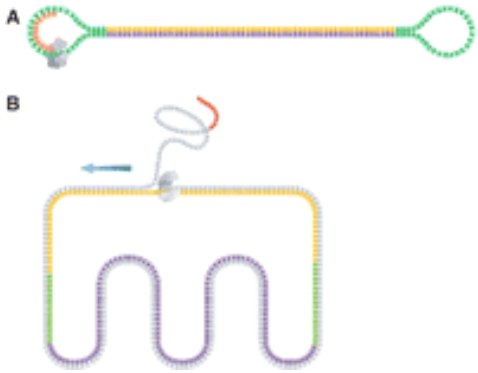
Helicos: true SMS

- First commercialized single molecule sequencer
- TIRF microscopy of sequencing by synthesis
- Readlengths of 25bp
- First de-commercialized single molecule sequencer
  - Just kidding. Mostly.

PacBio: single-molecule real-time

- Video of polymerases incorporating labeled nucleotides
- Optical confinement achieved with zero-mode waveguides
- Single molecule raw error rates higher than ensemble calls
- Unique template configuration
  - "SMRTbell™ template"
  - Enables circular consensus
- Readlengths …





Future of the site

- Q&A format (?)
  - Less personal but more information rich
  - http://biostar.stackexchange.com (1300 users and growing)
- Wiki expansion
- Recruit bloggers
- Conference
- Community science

- Find help
  - Designer
  - Wiki guru
  - PHP/Django/CMS/backend
  - Legal – official policy for information sharing
  - Opinionated bloggers

Acknowledgements

- Every member
- Volunteer moderators
  - Nils Homer and Lex Nederbragt
- Wiki Contributors, Dan Bolser
- Sponsors
- PacBio for being supportive
- Family…late nights and weekends spent on interweb