



UNIVERSITY
OF MANITOBA

AFS Seminar Series Fall 2013
October 23, 2013

Moving targets and millions of sequencing reads - Bioinformatics lessons from an international biofuels project

Dr. Brian Fristensky
Department of Plant Science
University of Manitoba
Winnipeg, Canada



GenomeCanada



MGCB² Microbial Genomics for Biofuels and
Co-Products from Biorefining Processes



GenomePrairie

Rationale

Bioinformatics - The science of management and analysis biological information, leading to biological discovery.

Bioinformatics

- **best done in the context of biological research**
- **organizes and drives research**
- **attempts to bridge the vast cultural and intellectual divides between biology and the computer and mathematical sciences**

Outline

- Original conception of the MGCB2 project
- Lessons learned
 - Organizing a world-wide genomics project
 - Genome assembly
 - Genome annotation - an endless game of tag
 - Gene expression - a moving target
 - Getting lost in the pathways
 - The cultural and intellectual divide between biologists and bioinformaticians
- The Bio Information Technologies Lab (Bit)

Microbial genomics - MGCB²

Microbial Genomics for Biofuels and Co-Products from Biorefining Processes

PIs: Dr. David Levin, Dr. Richard Sparling

Co-PIs

Trevor Charles (University of Waterloo)

Stephen Fong (Virginia Commonwealth Univ.)

Brian Fristensky (University of Manitoba)

Daniel Gapes (SCION, New Zealand)

Oleg Krokhin (University of Manitoba)

Kesen Ma (University of Waterloo)

Pin-Ching Maness (Natl. Renewable Energy Lab, Golden, Colorado)

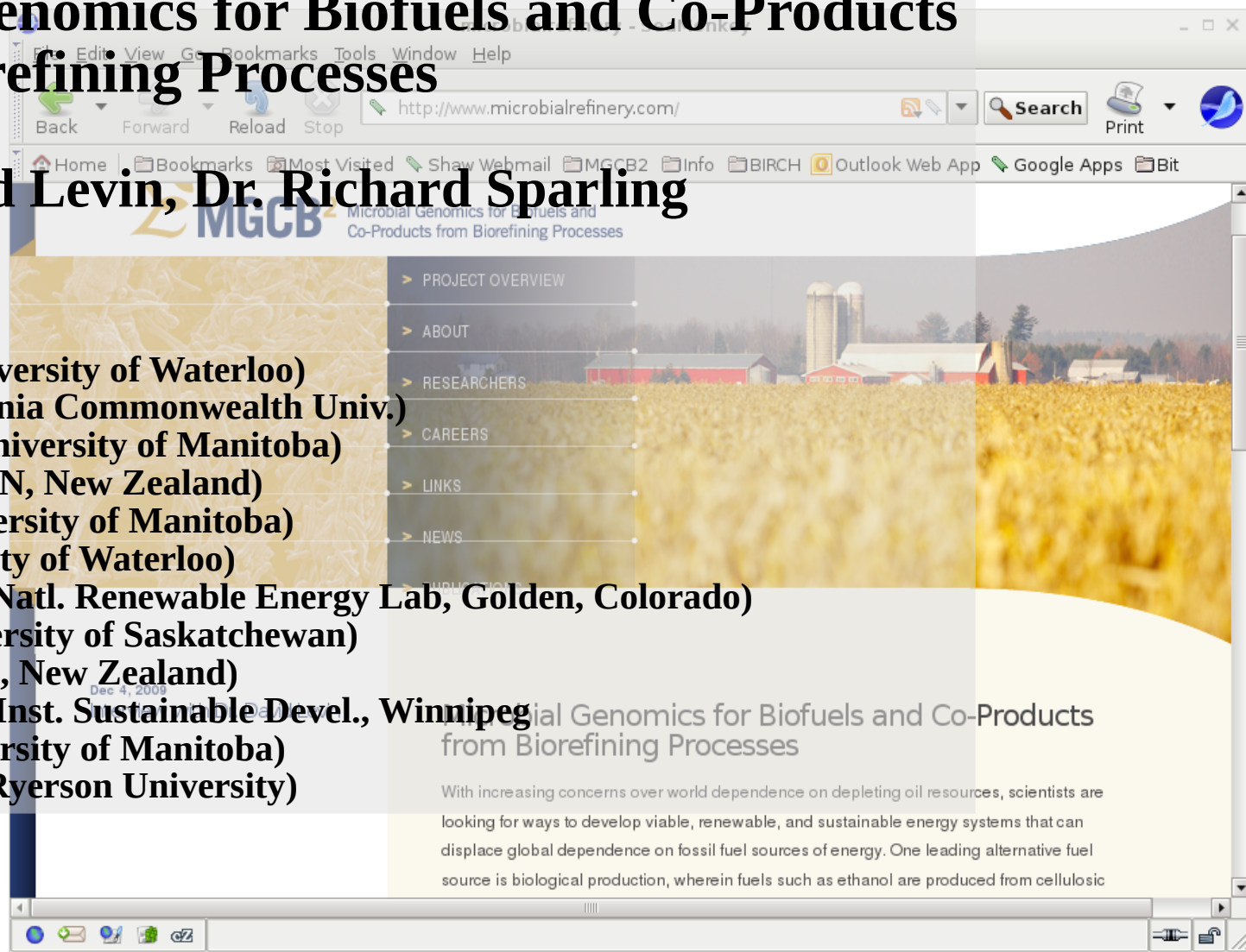
Stuart Smyth (University of Saskatchewan)

Matthew Stott (GNS, New Zealand)

Hank Venema, Intl. Inst. Sustainable Devel., Winnipeg

John Wilkins (University of Manitoba)

Gideon Wolfaardt (Ryerson University)



Microbial genomics - MGCB²

Goal: Improve efficiency of producing biofuels (eg. Ethanol, Butanol, H₂) from lignocellulosic feedstocks produced in agriculture (eg. biomass from wheat, flax, hemp etc.)

Strategy: Identify or engineer bacteria or bacterial consortia that can

- degrade cellulosic material into simple sugars, and
- ferment sugars into biofuels.

These organisms would be used in industrial-scale bioreactors to produce biofuels and related co-products, such as acetate, which can be used to create biodegradable plastics.

Microbial genomics - MGCB²

Activities:

Activity 1: Metabolic profiling and genomic characterization of bacteria that can utilize lignocellulosic biomass as a sole carbon source for synthesis of biofuels, as well as bacteria that can utilize the by-products of these fermentation reactions (sugars and/or organic acids) to synthesize polyhydroxylalkanoate (PHA) biopolymers;

Activity 2: Development of enhanced proteomic screening tools for rapid identification of the metabolic pathways used by bacteria, enabling selection of bacteria with the ability to convert cellulose to biofuels and/or cellulose hydrolysis products to PHAs;

Activity 3: Bioprospecting for novel cellulolytic and biopolymer synthesizing bacteria, including isolation and characterization at the genome level and determining suitability for their intended purpose within the biorefinery;

Activity 4: Metabolic engineering of selected bacteria to enhance synthesis of desired products;

Activity 5: Development of “plurifunctional designer consortia” for industrial application that can work together synergistically to enhance substrate conversion and product synthesis; and

Activity 6: An examination of GE³LS issues relevant to the project.

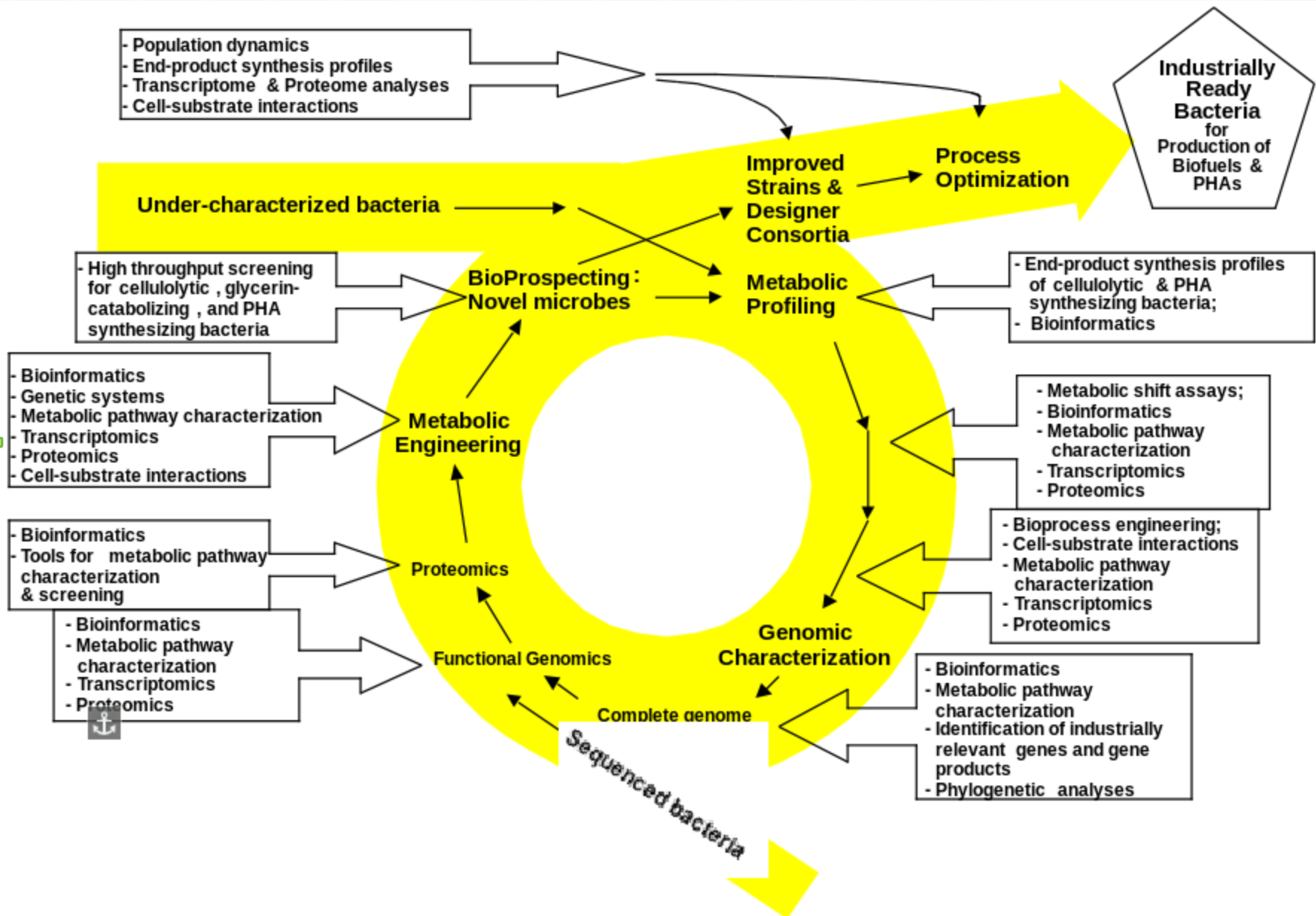
Lessons learned - The central theme:

The Shifty Paradigm:

Bioinformatics is a
Moving Target!



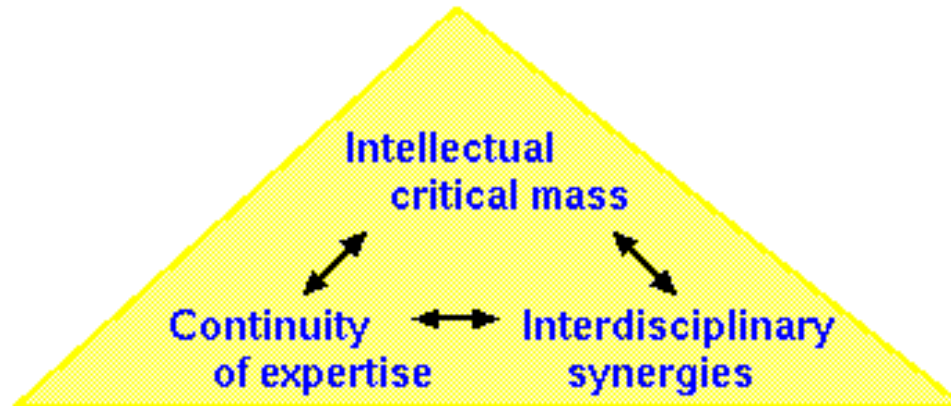
Organizing a world-wide genomics project



Bioinformatics Team, 2007 - 2013



UNIVERSITY
OF MANITOBA



Faculty
Brian Fristensky

Bioinformaticians
Graham Alvare
Justin Zhang
Maryam Ayat

PhD Student
Abiel Roche

Research Associate
Natalie Bjorklund
Ruming Li

Undergraduate Students
Dale Hamel
Drexler Hernandez

Organizing a world-wide genomics project

Location: Dept. of Computer Science

SunRay Thin clients

Mac OSX workstation/server (8 CPUs)

Linux x86_64 virtualization server (24 CPUs)

- Windows-XP
- Windows7
- Linux 64 -bit
- Linux 32-bit

Linux x86_64 test workstation/server (8 CPUs)

Xerox Phaser color printer

4 Tb offsite backup



Organizing a world-wide genomics project

CCU - Central Computing Unix

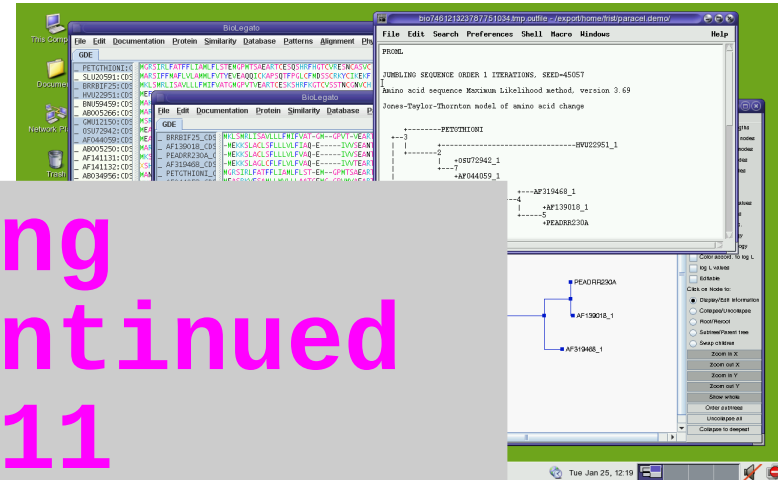
- Professionally-managed data centre
- Solaris, Linux, Windows
- Login servers (desktop sessions)
- Compute servers (CPU - intensive jobs)
- Open area labs
- Unified RAID filesystems
- Automated offsite backup

Organizing a world-wide genomics project

Genome Canada Bioinformatics Innovation Centre (BIC) Univ. Calgary

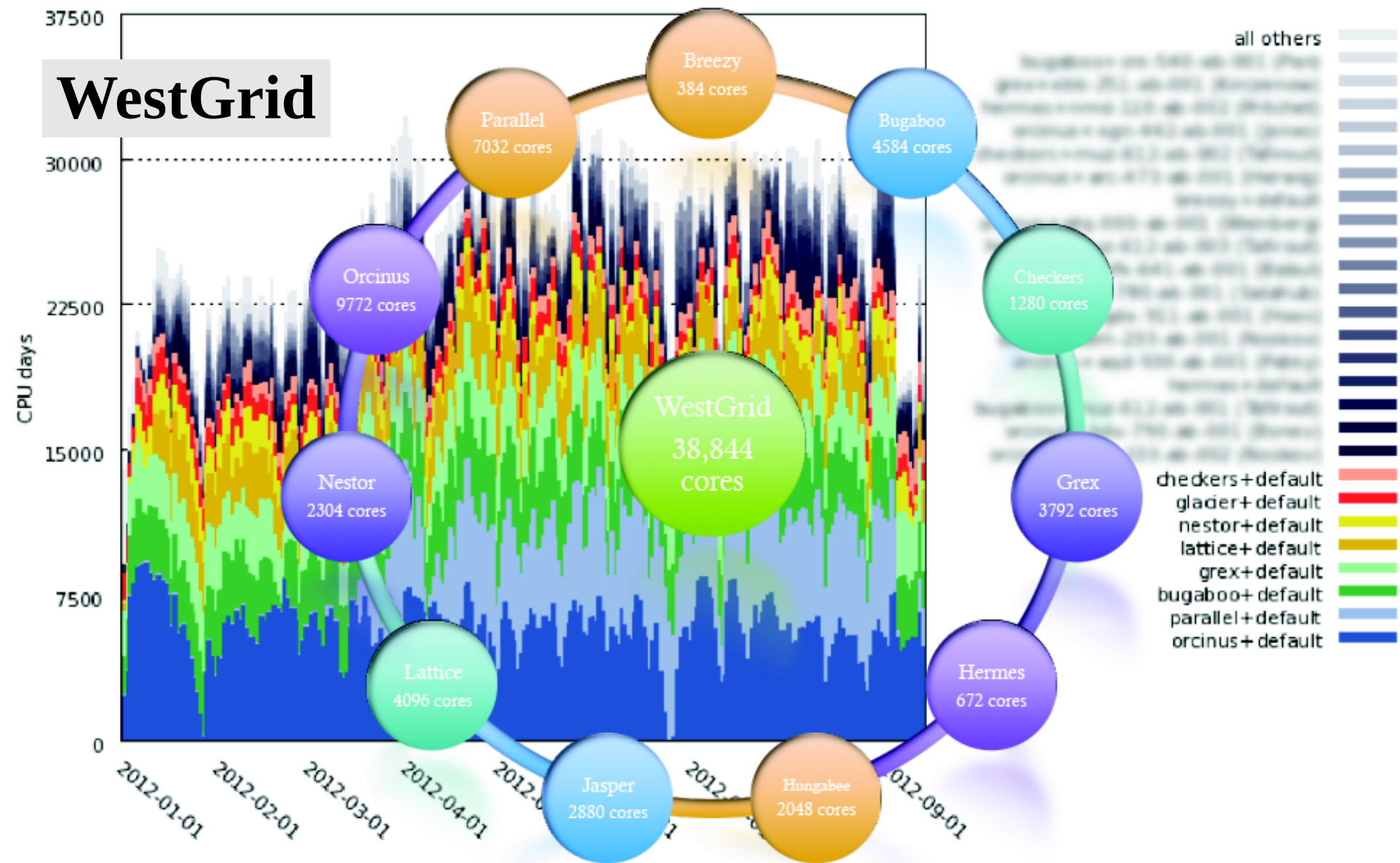
- Remote desktop session
- Hardware-accelerated database searches (TimeLogic)
- Data pipelines:
- MAGPIE - genome assembly and annotation
- Osprey - microarray design
- Phoenix - metagenomics

Funding discontinued in 2011 after 8 years



Organizing a world-wide genomics project

WestGrid

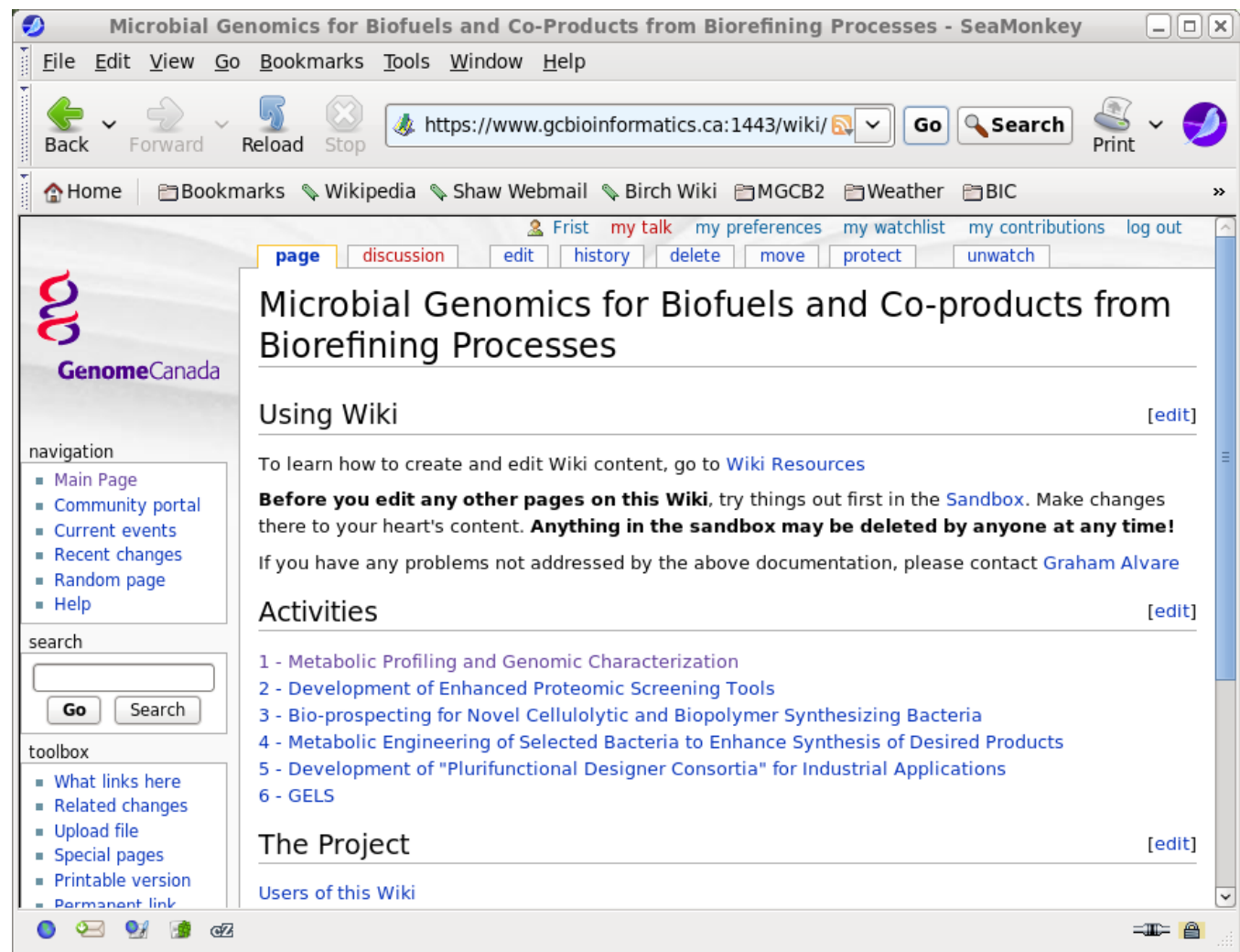


Organizing a world-wide genomics project

Wiki:
**Project-wide
collaboration**

**A Wiki is a
web site that
can be edited
by a group of
users.**

**Organizes
ideas,
documents,
data as the
project
proceeds.**



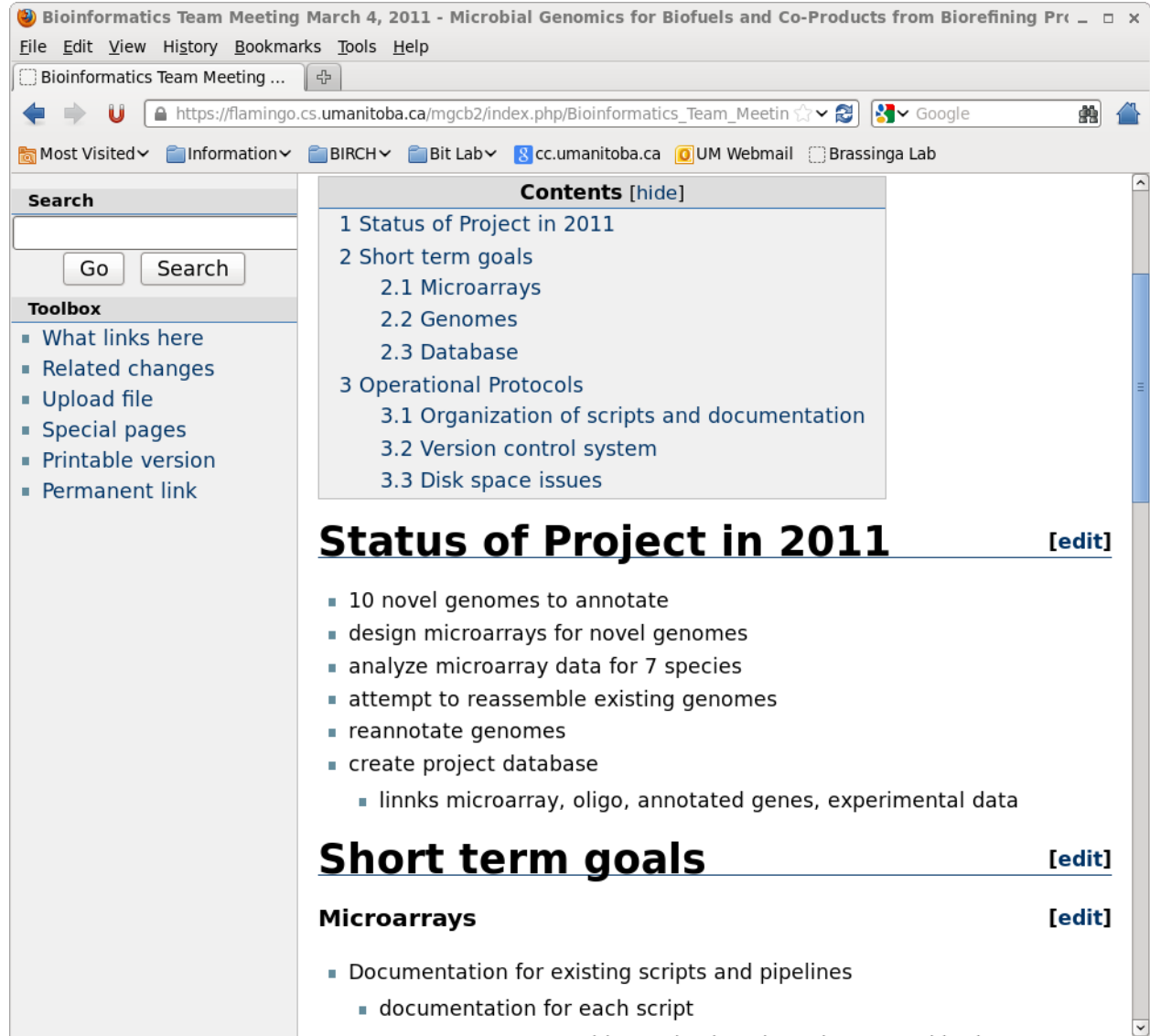
Organizing a world-wide genomics project

Meetings on the Wiki:

1) For each meeting, create an agenda in outline form.

2) During the meeting, fill in details as they are discussed.

The agenda becomes the minutes.



Bioinformatics Team Meeting March 4, 2011 - Microbial Genomics for Biofuels and Co-Products from Biorefining Pr...

File Edit View History Bookmarks Tools Help

Bioinformatics Team Meeting ...

https://flamingo.cs.umanitoba.ca/mgcb2/index.php/Bioinformatics_Team_Meetin

Most Visited Information BIRCH Bit Lab cc.umanitoba.ca UM Webmail Brassinga Lab

Search

Go Search

Toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

Contents [hide]

- 1 Status of Project in 2011
- 2 Short term goals
 - 2.1 Microarrays
 - 2.2 Genomes
 - 2.3 Database
- 3 Operational Protocols
 - 3.1 Organization of scripts and documentation
 - 3.2 Version control system
 - 3.3 Disk space issues

Status of Project in 2011 [edit]

- 10 novel genomes to annotate
- design microarrays for novel genomes
- analyze microarray data for 7 species
- attempt to reassemble existing genomes
- reannotate genomes
- create project database
 - linnks microarray, oligo, annotated genes, experimental data

Short term goals [edit]

Microarrays [edit]

- Documentation for existing scripts and pipelines
 - documentation for each script

Organizing a world-wide genomics project

*At the end of the project:
The Wiki is an archive data and ideas.*

Exponential cellobiose replicates (iTraq variations)

[edit]

Date: 2011-12-12

Experimenter: John Schellenberg

Phases: exponential

Metadata: Exponential cellobiose replicates. Samples A, B and C are biological replicates but sample C was prepared for analysis on a different day than A and B.

Full Metadata

- **Raw - MGF:** <https://flamingo.cs.umanitoba.ca/archive/mgcb2/csterc/raw/proteomic/csterc2Dnov9.txt>🔒
- **Intermediate - plotme:**
<https://flamingo.cs.umanitoba.ca/archive/mgcb2/csterc/intermed/proteomic/csterc2Dnov9-plotme.txt>🔒
- **Intermediate - plotme-NR:**
<https://flamingo.cs.umanitoba.ca/archive/mgcb2/csterc/intermed/proteomic/csterc2Dnov9-plotme-NR.txt>🔒

J0 - sample A prep 2 / sample A prep 1

[edit]

The screenshot shows a web browser window displaying the 'Archive/C.stercorarium' wiki page. The browser's address bar shows the URL 'https://flamingo.cs.umanitoba.ca/mgcb2/index.php/Archive/C.stercorarium'. The page has a blue header with the title 'Archive/C.stercorarium' and navigation links like 'FRIST', 'MY TALK', 'MY PREFERENCES', 'MY WATCHLIST', 'MY CONTRIBUTIONS', and 'LOG OUT'. A left sidebar contains a 'Navigation' menu with links to 'Main Page', 'Community portal', 'Current events', 'Recent changes', 'Random page', and 'Help', as well as a 'Search' box and a 'Toolbox' with links like 'What links here', 'Related changes', 'Upload file', 'Special pages', 'Printable version', and 'Permanent link'. The main content area shows a 'Contents' table of contents with a '[hide]' link. The table lists sections under '1 Genomics' and '2 Proteomics'. Under '1 Genomics', there are sub-sections 1.1, 1.2, and 1.3, each with further sub-sections. Under '2 Proteomics', there is sub-section 2.1, which has further sub-sections. The bottom of the page shows a small logo for 'Biofuels and Biorefining Processes'.

Archive/C.stercorarium - Microbial Genomics for Biofuels and Co-Products from Biorefining Processes - _ x

File Edit View History Bookmarks Tools Help

Archive/C.stercorarium - Microb... +

https://flamingo.cs.umanitoba.ca/mgcb2/index.php/Archive/C.stercorarium Google

Most Visited Information BIRCH Bit Lab cc.umanitoba.ca UM Webmail Brassinga Lab

Archive/C.stercorarium

FRIST MY TALK MY PREFERENCES MY WATCHLIST MY CONTRIBUTIONS LOG OUT

Navigation

- Main Page
- Community portal
- Current events
- Recent changes
- Random page
- Help

Search

Go Search

Toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

Contents [hide]

- 1 Genomics
 - 1.1 DSM 8532
 - 1.1.1 IMG/ER files (January 24, 2013) - Newbler DSM8532
 - 1.1.2 IMG/ER files (January 24, 2013) - GenePRIMP DSM8532
 - 1.1.3 Biocyc Pathway Tools PGDB files (January 24, 2013) - DSM 8532
 - 1.1.4 Raw sequencing reads (DSM 8532)
 - 1.2 DSM9219
 - 1.2.1 Raw sequencing reads (DSM 9219)
 - 1.3 DSM 2910
 - 1.3.1 Raw sequencing reads (DSM 2910)
- 2 Proteomics
 - 2.1 Exponential cellobiose replicates (iTraq variations)
 - 2.1.1 J0 - sample A prep 2 / sample A prep 1
 - 2.1.2 J1 - sample B / sample A prep 1
 - 2.1.3 J2 - sample C / sample A prep 1
 - 2.1.4 J3 - sample B / sample A prep 2
 - 2.1.5 J4 - sample C / sample A prep 2
 - 2.1.6 J5 - sample C / sample B

Organizing a genomics project - Lessons learned

- Wikis are powerful organizing tools - and easy to use.
- Free Open Source Software (FOSS) - no cost, and usually better than commercial.
- Redundancy of resources is essential
- Leverage existing infrastructure wherever possible.
- Factor in time for things to go wrong!
- The cultural gap between biologists and bioinformaticians can only be bridged through long-term multidisciplinary training of bioinformaticians.
- Work on a day to day basis with biologists - they are smarter than you think. :-)

Genome assembly

GS De Novo Assembler

Project: MO2_Draft [Genomic] Ready

Location: /local/workspace01/zhangju/Putida_MO2/MO2_Draft

Overview	Project	Parameters	Result files	Alignment results	Flowgrams																																																																																																																																																																																																																										
<div style="margin-bottom: 5px;">Contig ▲</div> <table style="width: 100%; border-collapse: collapse;"> <thead> <tr style="background-color: #d0e0ff;"> <th style="text-align: left; padding: 2px 5px;">Bases</th> <th style="text-align: left; padding: 2px 5px;">Bases</th> </tr> </thead> <tbody> <tr style="background-color: #007bff; color: white;"> <td style="padding: 2px 5px;">contig00001</td> <td style="padding: 2px 5px;">15,119</td> </tr> <tr> <td style="padding: 2px 5px;">contig00002</td> <td style="padding: 2px 5px;">1,031</td> </tr> <tr> <td style="padding: 2px 5px;">contig00003</td> <td style="padding: 2px 5px;">7,762</td> </tr> <tr> <td style="padding: 2px 5px;">contig00004</td> <td style="padding: 2px 5px;">2,509</td> </tr> <tr> <td style="padding: 2px 5px;">contig00005</td> <td style="padding: 2px 5px;">1,465</td> </tr> <tr> <td style="padding: 2px 5px;">contig00006</td> <td style="padding: 2px 5px;">2,626</td> </tr> <tr> <td style="padding: 2px 5px;">contig00007</td> <td style="padding: 2px 5px;">11,436</td> </tr> <tr> <td style="padding: 2px 5px;">contig00008</td> <td style="padding: 2px 5px;">1,705</td> </tr> <tr> <td style="padding: 2px 5px;">contig00009</td> <td style="padding: 2px 5px;">8,707</td> </tr> <tr> <td style="padding: 2px 5px;">contig00010</td> <td style="padding: 2px 5px;">3,158</td> </tr> <tr> <td style="padding: 2px 5px;">contig00011</td> <td style="padding: 2px 5px;">611</td> </tr> <tr> <td style="padding: 2px 5px;">contig00012</td> <td style="padding: 2px 5px;">768</td> </tr> <tr> <td style="padding: 2px 5px;">contig00013</td> <td style="padding: 2px 5px;">9,438</td> </tr> <tr> <td style="padding: 2px 5px;">contig00014</td> <td style="padding: 2px 5px;">6,560</td> </tr> <tr> <td style="padding: 2px 5px;">contig00015</td> <td style="padding: 2px 5px;">3,717</td> </tr> <tr> <td style="padding: 2px 5px;">contig00016</td> <td style="padding: 2px 5px;">747</td> </tr> <tr> <td style="padding: 2px 5px;">contig00017</td> <td style="padding: 2px 5px;">2,949</td> </tr> <tr> <td style="padding: 2px 5px;">contig00018</td> <td style="padding: 2px 5px;">5,751</td> </tr> </tbody> </table> <div style="margin-top: 10px;"> base: read: val: </div>	Bases	Bases	contig00001	15,119	contig00002	1,031	contig00003	7,762	contig00004	2,509	contig00005	1,465	contig00006	2,626	contig00007	11,436	contig00008	1,705	contig00009	8,707	contig00010	3,158	contig00011	611	contig00012	768	contig00013	9,438	contig00014	6,560	contig00015	3,717	contig00016	747	contig00017	2,949	contig00018	5,751	<div style="margin-bottom: 10px;"> Contig: contig00001 - 15,119 bp.:15119 <div style="float: right;"> <input type="text"/> Go </div> </div> <div style="margin-bottom: 10px;"> <table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left; padding: 2px 5px;">Base</th> <th style="text-align: left; padding: 2px 5px;">left 1,977</th> <th style="text-align: center; padding: 2px 5px;">selected ---</th> <th style="text-align: left; padding: 2px 5px;">right 2,054</th> <th style="padding: 2px 5px;">-</th> <th style="padding: 2px 5px;"><div style="width: 50px; height: 10px; background: linear-gradient(to right, blue, white);"></div></th> <th style="padding: 2px 5px;">+</th> <th style="padding: 2px 5px;">AC</th> <th style="padding: 2px 5px;">G</th> </tr> </thead> <tbody> <tr> <td style="padding: 2px 5px;">contig00001</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302DRCSK_left</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302EDDD7_right</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302C6KML_left</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302ELTON</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302C5V68_left</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302D3CFV_right</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302DX0BN_left</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302DY70L</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302DSPME_left</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302DLW3S_right</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302EH0U6_left</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302ESP3D</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302SDX7_right</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302D1AC9_right</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302DNRMN_left</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302DZQP_right</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302EHJ6K_right</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> <tr> <td style="padding: 2px 5px;">H9Y0BK302COM39_right</td> <td colspan="8" style="padding: 2px 5px;">GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC</td> </tr> </tbody> </table> </div>					Base	left 1,977	selected ---	right 2,054	-	<div style="width: 50px; height: 10px; background: linear-gradient(to right, blue, white);"></div>	+	AC	G	contig00001	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302DRCSK_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302EDDD7_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302C6KML_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302ELTON	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302C5V68_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302D3CFV_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302DX0BN_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302DY70L	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302DSPME_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302DLW3S_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302EH0U6_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302ESP3D	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302SDX7_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302D1AC9_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302DNRMN_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302DZQP_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302EHJ6K_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC								H9Y0BK302COM39_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC							
Bases	Bases																																																																																																																																																																																																																														
contig00001	15,119																																																																																																																																																																																																																														
contig00002	1,031																																																																																																																																																																																																																														
contig00003	7,762																																																																																																																																																																																																																														
contig00004	2,509																																																																																																																																																																																																																														
contig00005	1,465																																																																																																																																																																																																																														
contig00006	2,626																																																																																																																																																																																																																														
contig00007	11,436																																																																																																																																																																																																																														
contig00008	1,705																																																																																																																																																																																																																														
contig00009	8,707																																																																																																																																																																																																																														
contig00010	3,158																																																																																																																																																																																																																														
contig00011	611																																																																																																																																																																																																																														
contig00012	768																																																																																																																																																																																																																														
contig00013	9,438																																																																																																																																																																																																																														
contig00014	6,560																																																																																																																																																																																																																														
contig00015	3,717																																																																																																																																																																																																																														
contig00016	747																																																																																																																																																																																																																														
contig00017	2,949																																																																																																																																																																																																																														
contig00018	5,751																																																																																																																																																																																																																														
Base	left 1,977	selected ---	right 2,054	-	<div style="width: 50px; height: 10px; background: linear-gradient(to right, blue, white);"></div>	+	AC	G																																																																																																																																																																																																																							
contig00001	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302DRCSK_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302EDDD7_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302C6KML_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302ELTON	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302C5V68_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302D3CFV_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302DX0BN_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302DY70L	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302DSPME_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302DLW3S_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302EH0U6_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302ESP3D	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302SDX7_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302D1AC9_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302DNRMN_left	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302DZQP_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302EHJ6K_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														
H9Y0BK302COM39_right	GTCGG-CCTCAAGCACAGC-AAGCCCCTGGGCAACTGGG-TGCTCAGTGCCAACCTGGGGTATTTACCGCA-CGCGAGGAC																																																																																																																																																																																																																														

Genome assembly - raw reads

Name	Alias	Number of Bases	Number of Reads 454	Number of Reads HiSeq
1635	Chloroflexus	163013840	542487	0
454-1863	Isolate WC-1	80293876	242971	0
454-1866	Caldicellulosiruptor kristjanssonii	72135175	232506	0
454-1866_2	Caldicellulosiruptor kristjanssonii	0	0	0
454-1920	Clostridium termitidis	109064533	303437	0
454-1921	Clostridium novel species	98307634	284382	0
454-1922	Geobacillus debilis	74722425	207025	0
85F	85F	0	0	0
Clostridium_stercorarium_DSM 8532	Clostridium stercorarium DSM8532	84294306	237698	0
Novel_clostridium	Novel_clostridium	34309169	92277	0
P.putida_LS46_replacement	P.putida_LS46_replacement	31042996000	0	155214980
Pseudomonas_putida_LS46	Pseudomonas putida LS46	85468673	230815	0
Strain_WC1	Strain_WC1	24408440600	0	122042203
Tg-Thermococcus	Tg-Thermococcus	128812328	400318	0
Th-Thermotoga_hypogea	Th-Thermotoga hypogea	133910103	352497	0

Genome assembly - Thermoanaerobacter twc1

No	Reads	Pipeline	Number of contigs	Longest (bp)	Mean (bp)	N50 (bp)	total bases (bp)
1	454	Newbler v2.3	123	182175	20815	48240	2560197
2	Illumina	VelvetOptimiser(insertion size 277 sd 35 from Ray result) option Longest	128	245148	20008	63283	2561085
3	Illumina	VelvetOptimiser (insertion size 330 from sequencing Bioanalyser) option Longest	150	113571	17165	48456	2574783
4	mix	VelvetOptimiser(insertion size 330) option Longest	371	227764	6882	20205	2553227
5	mix	VelvetOptimiser no insert size option N50	155	281930	16756	50951	2597148
6	Illumina	Ray No insert size input	347	73471	7449	13048	2584799
7	Illumina	Ray insert size 300 sd 60	341	73471	7570	14380	2581650
8	mix	Ray No insert size input	164	105499	15426	31846	2529922
9	mix	Ray insert size 300 sd 60	158	105546	16045	31864	2535250

Genome assembly - Lessons learned

- Sequencing technologies keep changing the game
- Software evolves along with seq. technology
- Evaluating software is a non-trivial problem
- Most genomes are $< 100\%$ complete
- There is no such thing as an off the shelf pipeline
- Each genome provides unique challenges
- The state of the art programs are a moving target

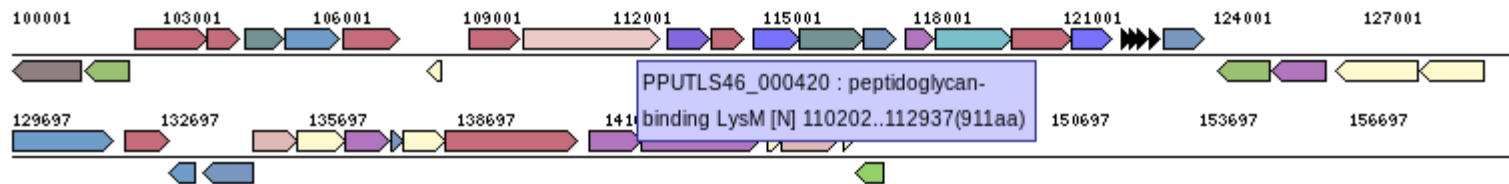
Genome annotation - an endless game of tag

Chromosome Viewer - Colored by COG

Switch coloring to:

[Pseudomonas putida LS46 \(NCBI Annotation V2\) : ALPV02000001](#) (212968bp gc=0.62 depth=1.00)
(coordinates **100001-150000**)

hint: Mouse over a gene to see details.
Query gene in **red**, RNAs in **black**, Pseudo genes in **white**
My Gene in **cyan** or dashes
Gene(s) with protein is marked by a **purple** bar
Gene(s) in Gene Cart is marked by a **blue** bar
|||||| CRISPR array



< Previous Range

Next Range >

[Get Nucleotide Sequence For Range](#)

External Links:
- taxon:1193146

Genome annotation - an endless game of tag

Example: Correspondence of gene tags between datasets (*C. thermocellum*)

JW20_3317	NZ_ABVG02000001.1	gene_symbol	579372..577336(-)	type 3a cellulose-binding domain protein
2360_5448	NZ_ACVX01000021.1	gene_symbol	26916..24880(-)	type 3a cellulose-binding domain protein
27405_0262	NC_009012.1	gene_symbol	331028..330231(-)	putative RNA polymerase sigma factor SigI putative RNA po
2360_5448	NZ_ACVX01000021.1	gene_symbol	27694..26897(-)	RNA polymerase, sigma 28 subunit, SigI
27405_0269	NC_009012.1	gene_symbol	341280..343715(+)	cellobiose phosphorylase glycosyltransferase 36
2360_5454	NZ_ACVX01000021.1	gene_symbol	37944..40379(+)	glycosyltransferase 36
27405_0270	NC_009012.1	gene_symbol	343980..344939(+)	D-isomer specific 2-hydroxyacid dehydrogenase NAD-binding
2360_5454	NZ_ACVX01000021.1	gene_symbol	40644..41603(+)	D-isomer specific 2-hydroxyacid dehydrogenase NAD-binding
JW20_3324	NZ_ABVG02000001.1	Cther_1654	595674..595561(-)	hypothetical protein
2360_5454	NZ_ACVX01000021.1	ClothDRAFT_1821	41820..41707(-)	hypothetical protein

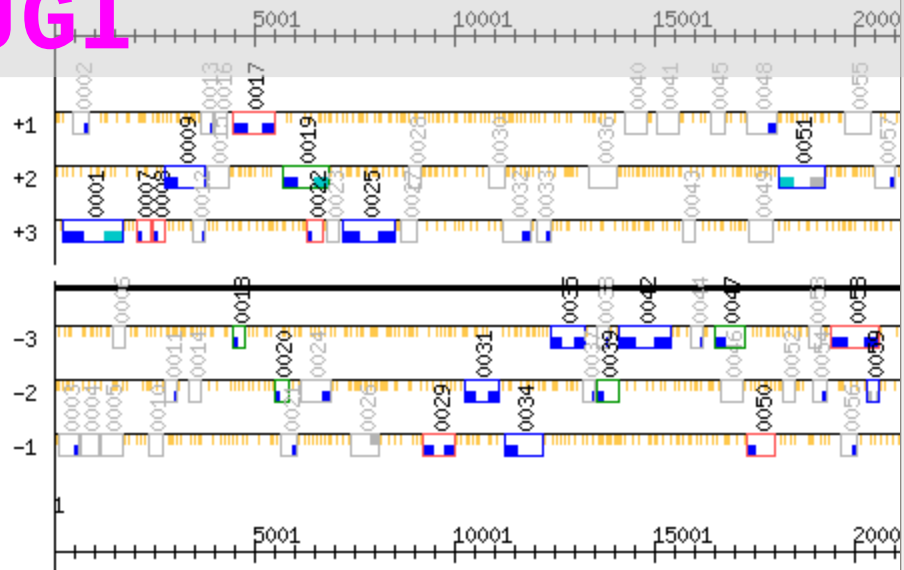
Genome annotation - an endless game of tag

Example: Resolving ambiguities and conflicts The same enzyme with 3 different EC numbers!

- Enzyme #1 acetaldehyde dehydrogenase
EC=1.2.1.10
- Enzyme #2 alcohol dehydrogenase EC=1.1.1.1
- Enzyme #3 acetaldehyde/alcohol
dehydrogenase EC=1.1.1.1 and 1.2.1.10
- If we link enzymes only to BRENDA we miss
#3

Genome annotation - MAGPIE annotation

Had to discard
when BIC was
discontinued
and start over at
JGI



Minimum length = 100 (AA residues) ORF traits:

f	a	c
p	d	m

contig00015_00054 function form - SeaMonkey

File Edit View Go Bookmarks Tools Window Help

contig00015_00054

/magpie/private/ck_glimmer/html/contig00015: No such file or directory

Annotation for ck_glimmer contig00015 contig00015_00054

There is no annotation.

MAGPIE Suggestion: **bacterioferritin** [B] last analysis 19/11/10
based on summary evidence (level 1) synthesis
contig00015_00054 sequence properties: 507 BP 507 putative CDS 507 putative protein

Possible ontology terms with confidence level 1 add # Update

Status	Lvl	EC	Gene Name	Gene Prod.	Description
Add putative	1				bacterioferritin
Start Codon Pos(AA)	1	Stop	507	Password	Seq Alias

SUPPORTING EVIDENCE:

1 67 134 169
Start codons ATG GTG TTG (Short tick=rare codon)

contig00015_00054

Protein similarity

DNA similarity

Motif

123 interpro

123 prosite

Genome annotation - an endless game of tag

➤ **PRODIGAL** - prokaryotic gene prediction

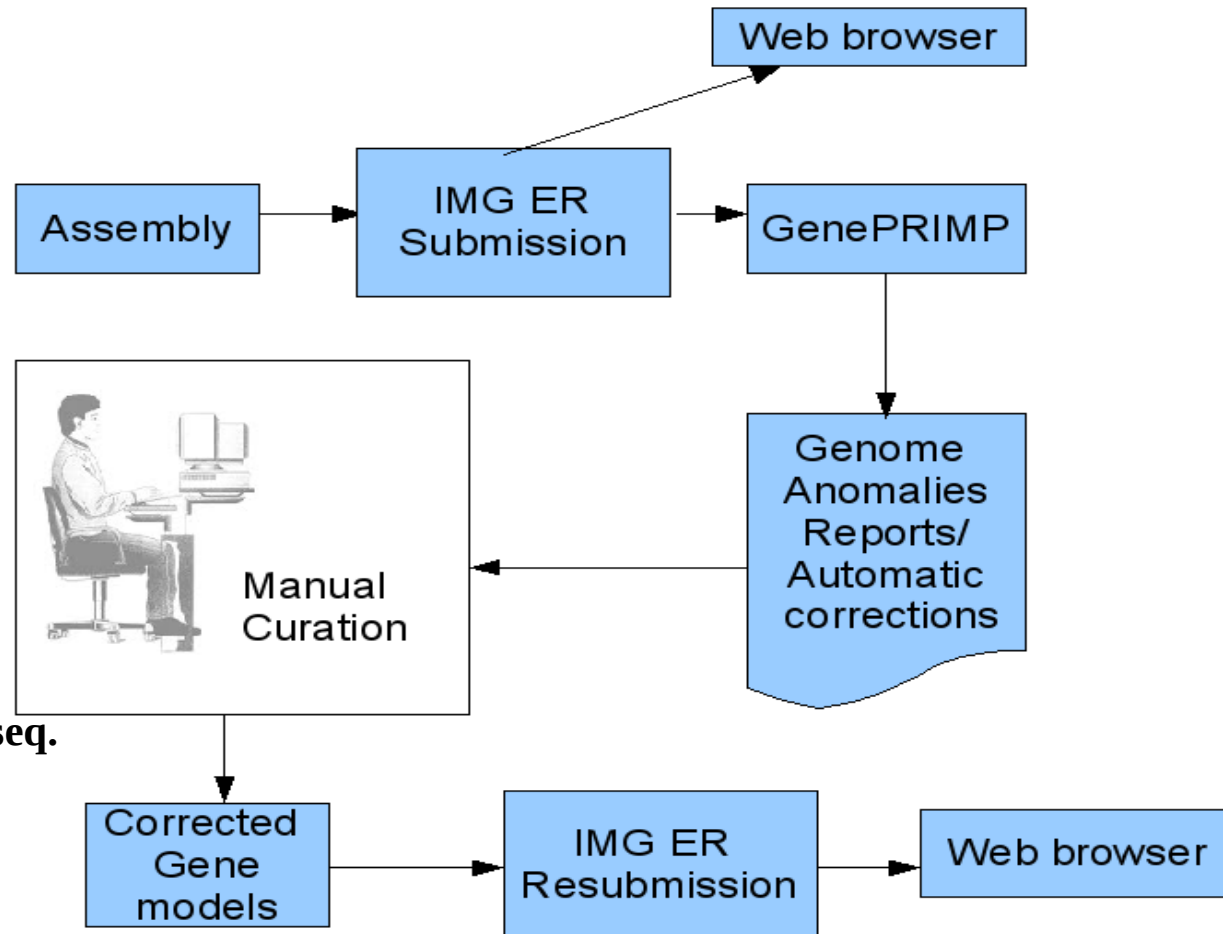
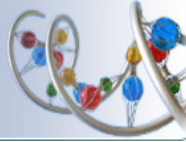
➤ **Low false positive
rate; performs well
even for genomes of
high GC %**

➤ **GenePRIMP** – Gene Prediction Improvement Pipeline

- **short** - 5' truncated
- **long** - 3' truncated
- **unique** - no known homologues
- **dubious** - too short
- **split** - interrupted by frameshifts
- **split** - interrupted by transposons
- **missed** - homologous to database seq.
- **CRISPR** - clustered short palin-
- **dromic repeats**



DOE Joint Genome Institute
Enabling Advances in Bioenergy & Environmental Research



GenomeCanada

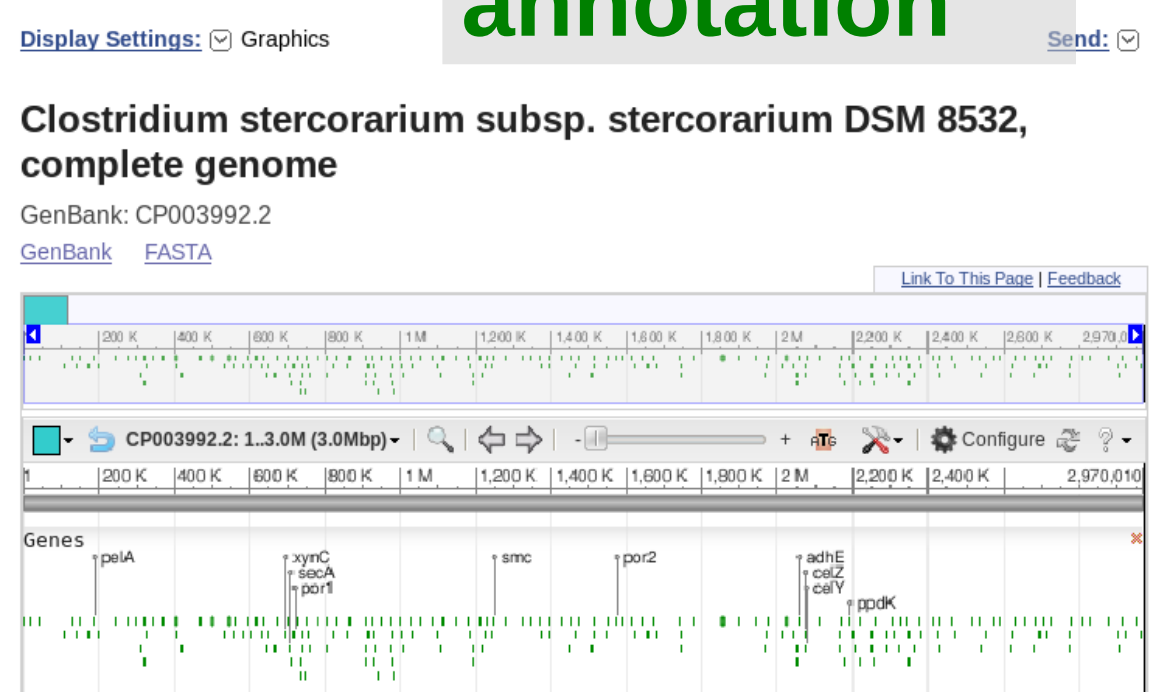
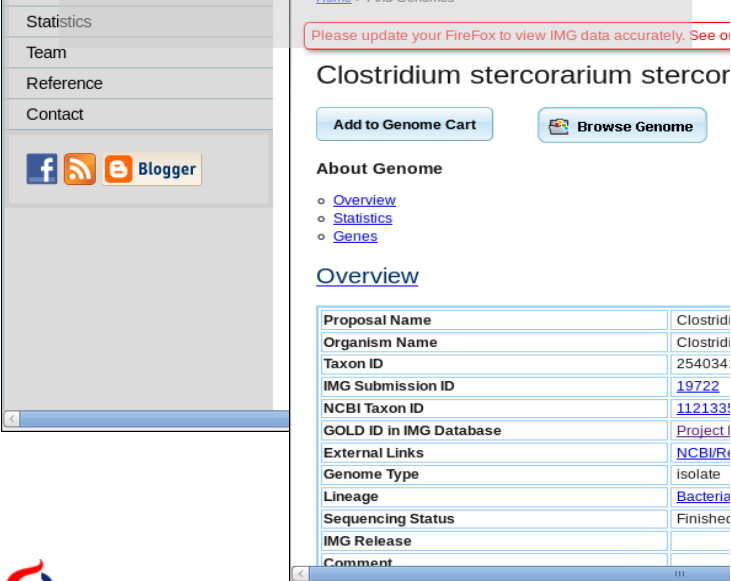
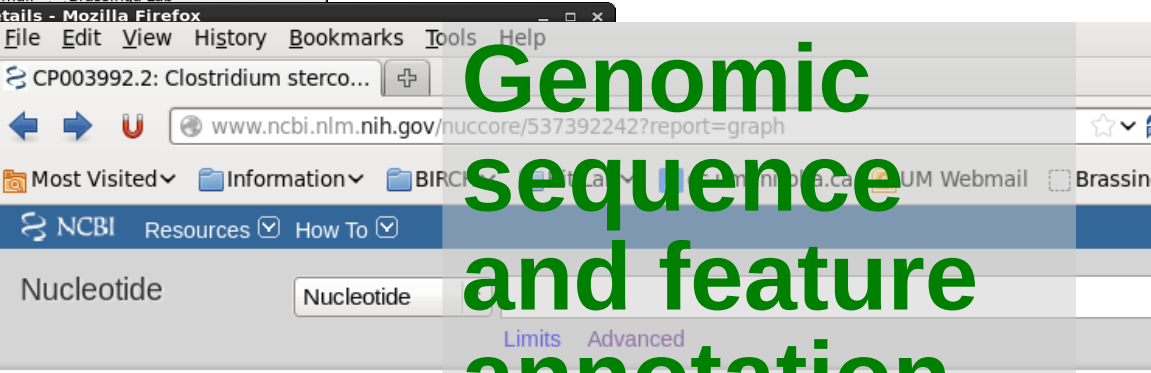
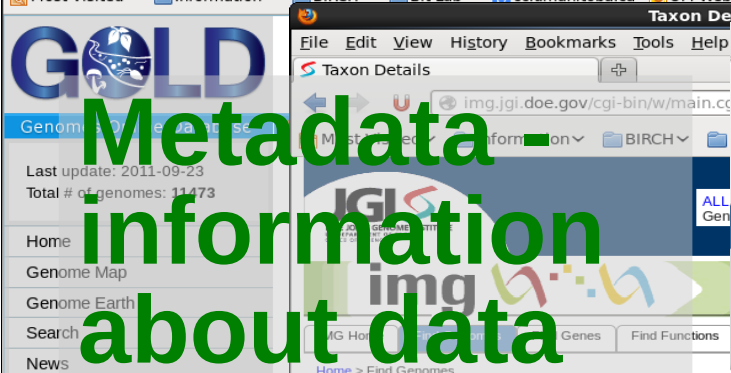
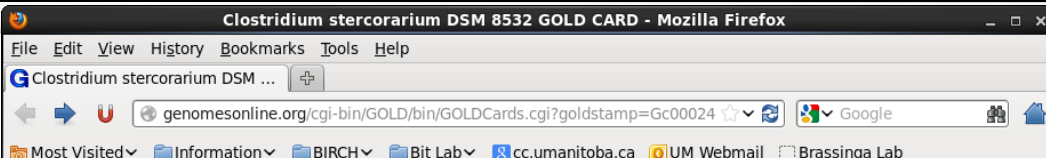
MGCB² Microbial Genomics for Biofuels and
Co-Products from Biorefining Processes

GenomePrairie

Genome annotation - an endless game of tag

Metadate - information about data

Genomic sequence and feature annotation



Genome annotation - Lessons learned

- **Not only is there no off the shelf pipeline, but you better be prepared to write your own scripts to fill in the gaps, between what one program does and the next program**
- **Gene tags NEVER agree between different databases**
- **The trick is to automate as much as possible, and have the computer flag those genes that need expert human judgement to complete the annotation.**

Gene expression

University_of_Manitoba-MGCB2_Pilot_Study_C_thermocellum_Microarray_Experiment.

File Edit View Insert Format Tools Data Window Help

Arial 10

A1 $f(x)$ Σ =

	A	B	C	D	E	F
1		ProbeName	gProcessedSignal_stdev	numOfProbes	gProcessedSignal_mean	gProcessedSignal_Cv
2	1	(-)3xSLv1	359.3696168165	77	90.0248521558	3.9918934407
3	2	(+)E1A_r60_1	27613.536855661	20	307954.1	0.0896677033
4	3	(+)E1A_r60_3	36.424649619	20	35.1165026	1.0372516316
5	4	(+)E1A_r60_a104	94.8346982573	20	327.433385	0.2896305099
6	5	(+)E1A_r60_a107	71.126314748	20	413.8407	0.1718688248
7	6	(+)E1A_r60_a135	584.0851887314	20	3454.45255	0.1690818387
8	7	(+)E1A_r60_a20	408.6686393613	20	1325.1875	0.3083855223
9	8	(+)E1A_r60_a22	471.640633383	20	3597.54725	0.1311006084
10	9	(+)E1A_r60_a97	2307.529616655	20	26322.3315	0.0876643323
11	10	(+)E1A_r60_n11	5407.1545814228	20	59141.327	0.0914276844
12	11	(+)E1A_r60_n9	17069.6308911038	20	190345.98	0.0896768657
13	12	(+)eQC-39	86.5837520929	6	55.8165776667	1.5512192921
14	13	(+)eQC-40	32.1026883533	6	25.748531	1.2467774707
15	14	(+)eQC-41	13.4440872145	6	15.39848	0.8730788503
16	15	(+)eQC-42	116.2240343315	6	187.4691666667	0.619963466
17	16	2360_4736	1491.3585186736	3	5479.0466666667	0.2721930674
18	17	2360_4747	2521.2553807892	4	21316.1675	0.118279019
19	18	2360_4755	2082.0088817741	3	11170.2176666667	0.1863892848
20	19	US09503746_253417410001_S01_GE1_1010_Sep10_2_4				
21	20	US09503746_253417410002_S01_GE1_1010_Sep10_2_2				

Find

Sheet 11 / 26 PageStyle_US09503746_253417410001_S01_GE1_1010_Sep10_2_2 STD Sum=

Gene expression - custom microarrays

For new species, you can't just buy arrays from Agilent! You must design oligos from ORFs

A further complication: We want to include ORFs from 3 *C. thermocellum* strains:

- many redundant ORFs
- gene tags different between strains

Gene expression - custom microarrays

Example: some ORFs are identical except for the start or stop codon.

Table 3. Five pairs of aligned sequences with the same base compositions except start or stop codon.

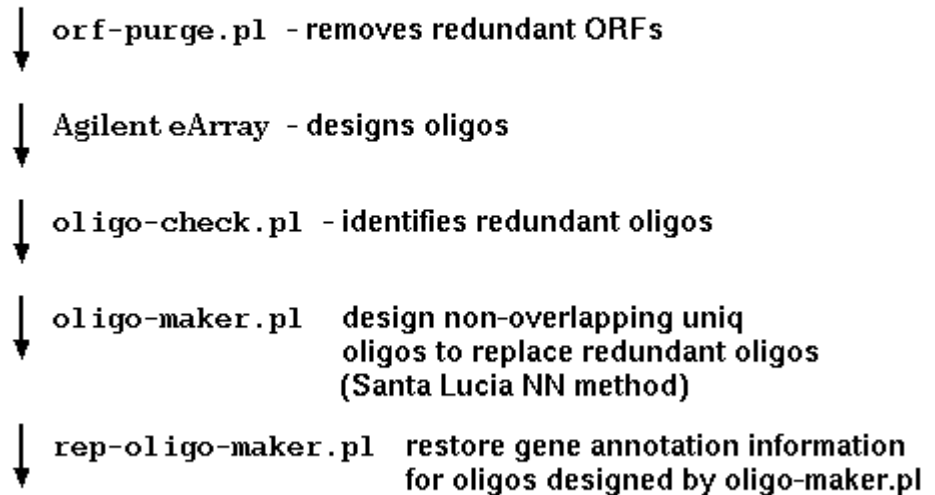
Gene_ID	Gene sequences with identical ORF bodies
27405_1409	TTG AAAACGTGTATTGCTTGTGGAATGCCTATGAAGGATATTCAGACTTT//TATTTTA ATTAA
2360_5735	ATG AAAACGTGTATTGCTTGTGGAATGCCTATGAAGGATATTCAGACTTT//TATTTTA ATTAA
JW20_4146	ATG GCAGGGGATCATTATATGCTCCGTTTGTGGAAAGTGTAAGGAAAGTT//ACTGAC AAGTAA
2360_5047	ATG GCAGGGGATCATTATATGCTCCGTTTGTGGAAAGTGTAAGGAAAGTT//ACTGAC AAGTAG
27405_2041	ATG AGCTTTTATGCCGCACCGATTGCAAGGCTTATAGAGGAGTTTGAGAAAG//CGCGA GATTAG
JW20_4452	ATG AGCTTTTATGCCGCACCGATTGCAAGGCTTATAGAGGAGTTTGAGAAAG//CGCGA GATTAA
27405_2221	ATG AGCGCGAAAATCCTTGTTGTTGATGACGAGAAAATATAGTTGACATT//AAATTA AGTTAA
JW20_4195	GTG AGCGCGAAAATCCTTGTTGTTGATGACGAGAAAATATAGTTGACATT//AAATTA AGTTAA
27405_2542	TTG TGGGTATCTGTTAGCAATCAGGCATATGTTTTTTTAAATTGTGTTCTC//AAAAAAAT ATAA
JW20_4049	TTG TGGGTATCTGTTAGCAATCAGGCATATGTTTTTTTAAATTGTGTTCTC//AAAAAAAT ATAG

Gene expression - custom microarrays

Design one inclusive array for ORFs from 3 strains

Clostridium thermocellum

ATCC 27405 (DSM 1237)
JW20 (DSM 4135)
DSM 2360



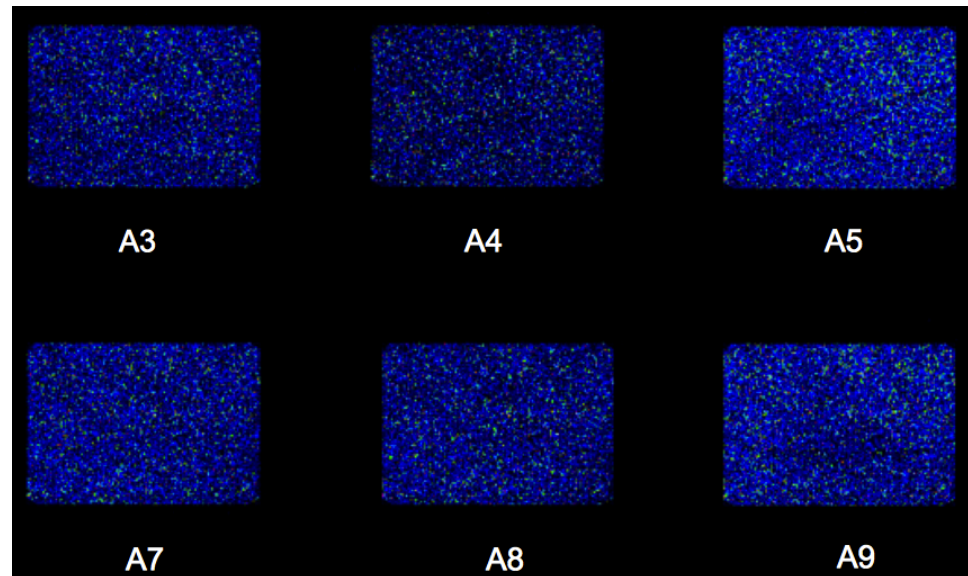
60-mer oligos

Gene expression - a moving target

Pilot Experiment - Assessment of statistical power by resampling of biological replicates

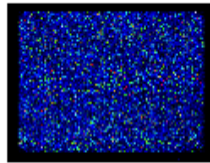
300 genes
8 duplicate spots

3597 genes
3 duplicate spots

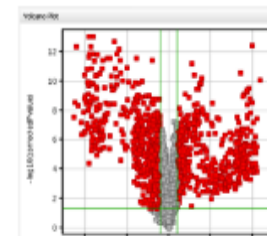
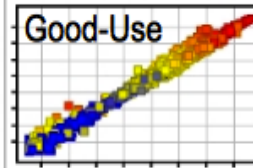
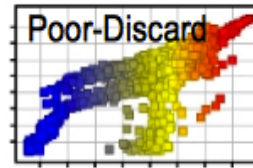


Gene expression - a moving target

2. Quality Control Checks



Images Manually
Checked

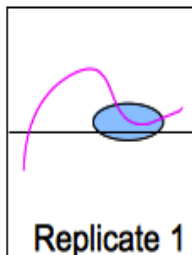


Correlation plots and volcano plots check cross sample technical quality

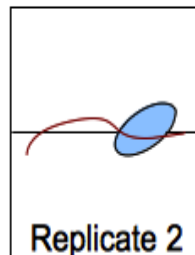
Agilent Feature Extraction software

3. Normalization

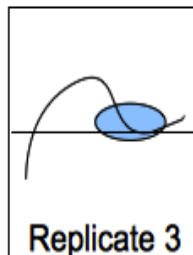
(removes variation due to non-biological factors between arrays)



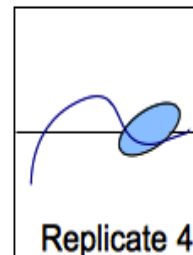
Replicate 1



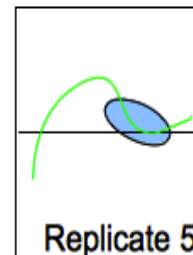
Replicate 2



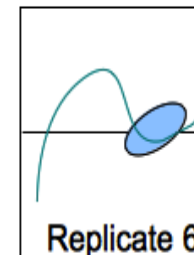
Replicate 3



Replicate 4

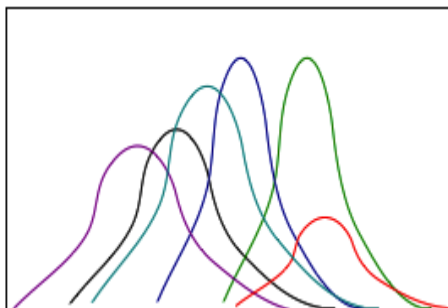


Replicate 5

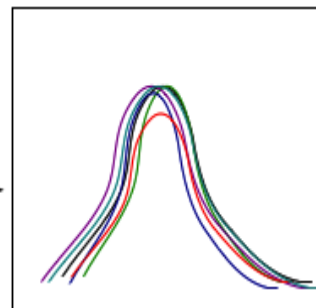


Replicate 6

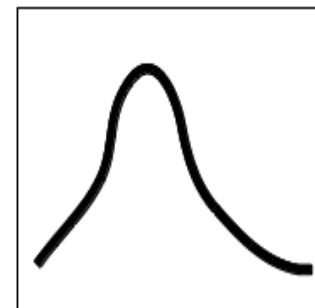
First test if we need to normalize using MA plots – Yes?



Raw Data



Normalized Data



Averaged over replicates if required

Gene expression - a moving target

How many biological replicates are needed?

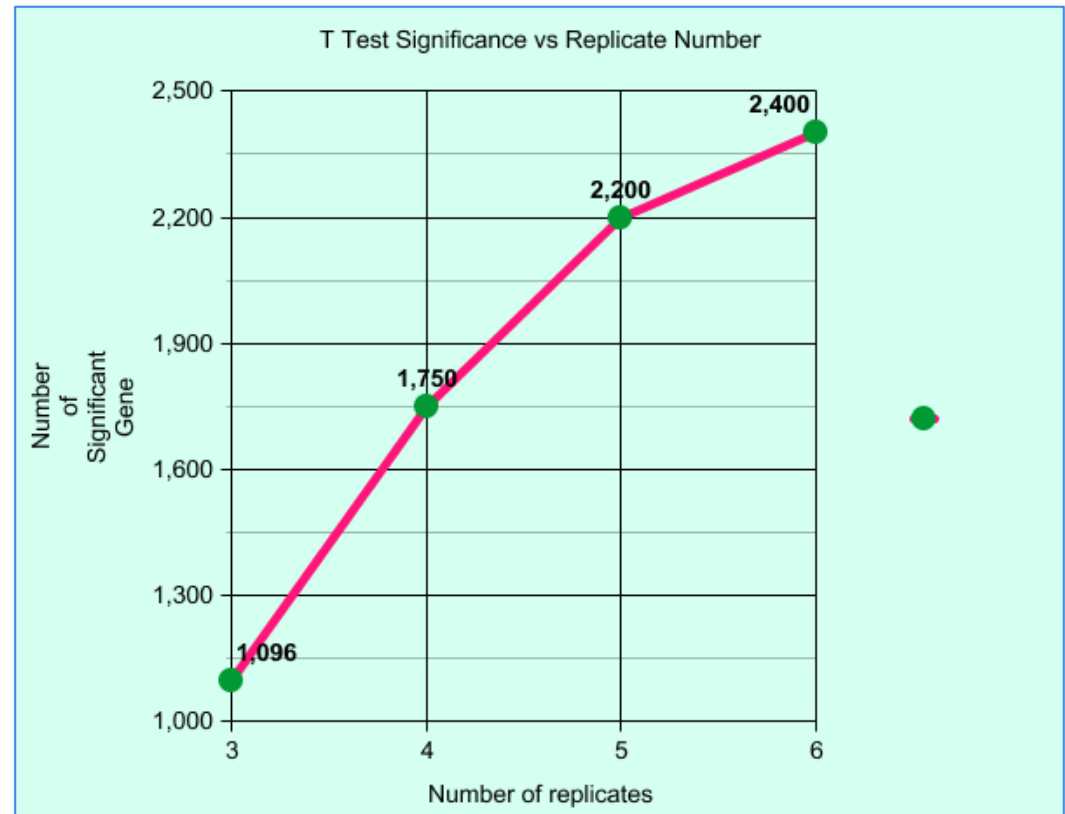
**Normalized Gene
Microarray 4233 probes**

**A - set of 6 Clostridium
thermocellum 1237
exponential**

**B - set of 6 Clostridium
thermocellum 2360
exponential**

**C - set of 5 Clostridium
thermocellum 4150 exp
(discarded)**

**D - set of 6 Clostridium
thermocellum 1237
stationary**



Gene expression - a moving target

RNA sequencing

1. > 100 million reads per sample
2. count # of reads for each gene
3. correct for different sizes of genes

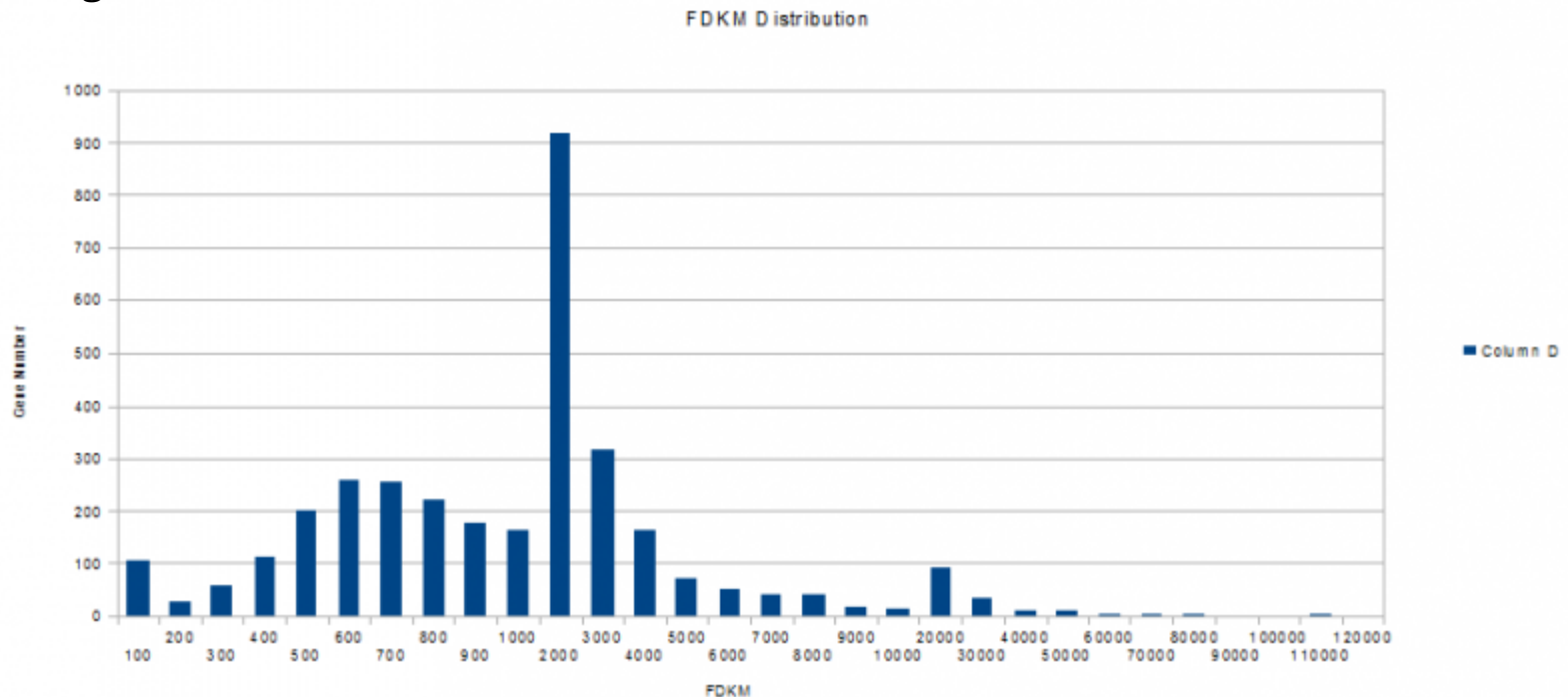
Intrinsically simpler than microarrays, so fewer sources of experimental variance.

Gene expression - a moving target

Illumina RNA Seq of *C. thermocellum*

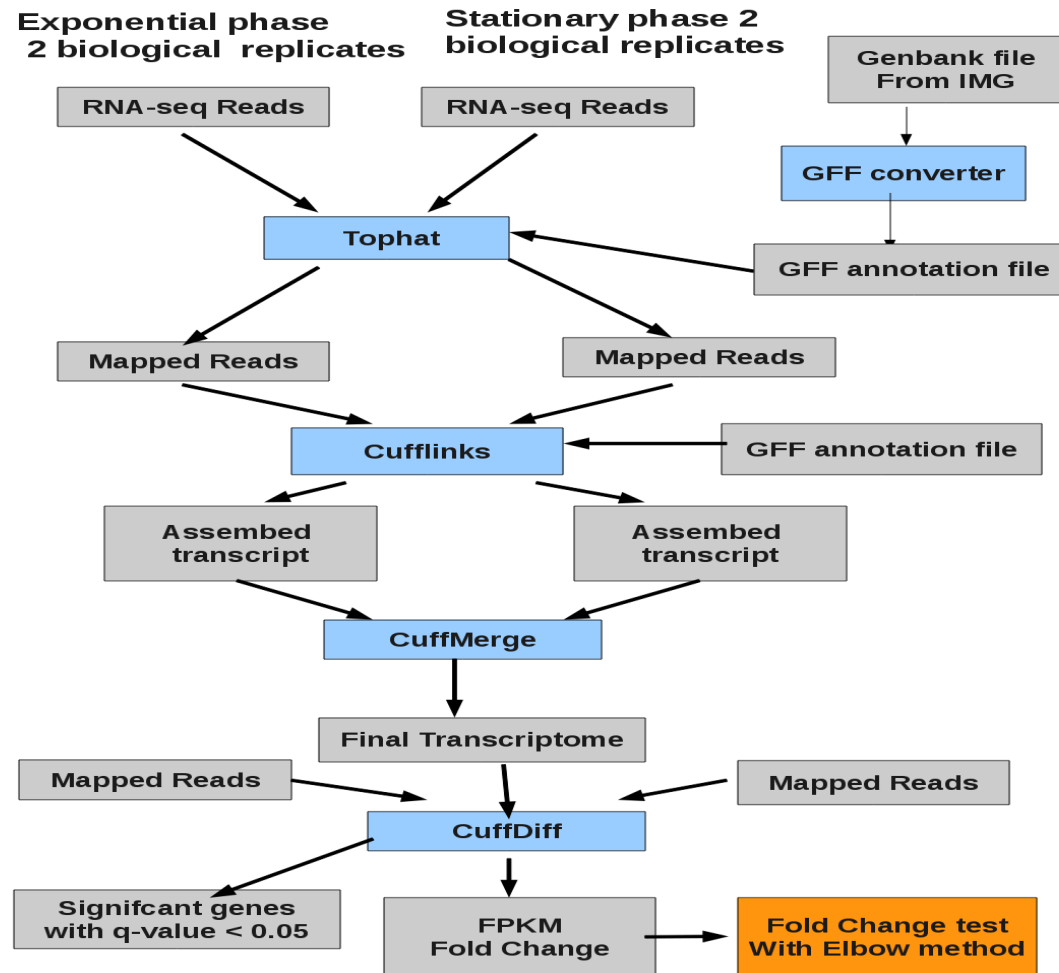
152 milion reads
95% at least 1 alignment
5% failed to align

FPKM - fragments per kilobase of transcript per
million fragments mapped



Pipeline: Tophat, Bowtie, Cufflink

Gene expression - RNAseq data pipeline



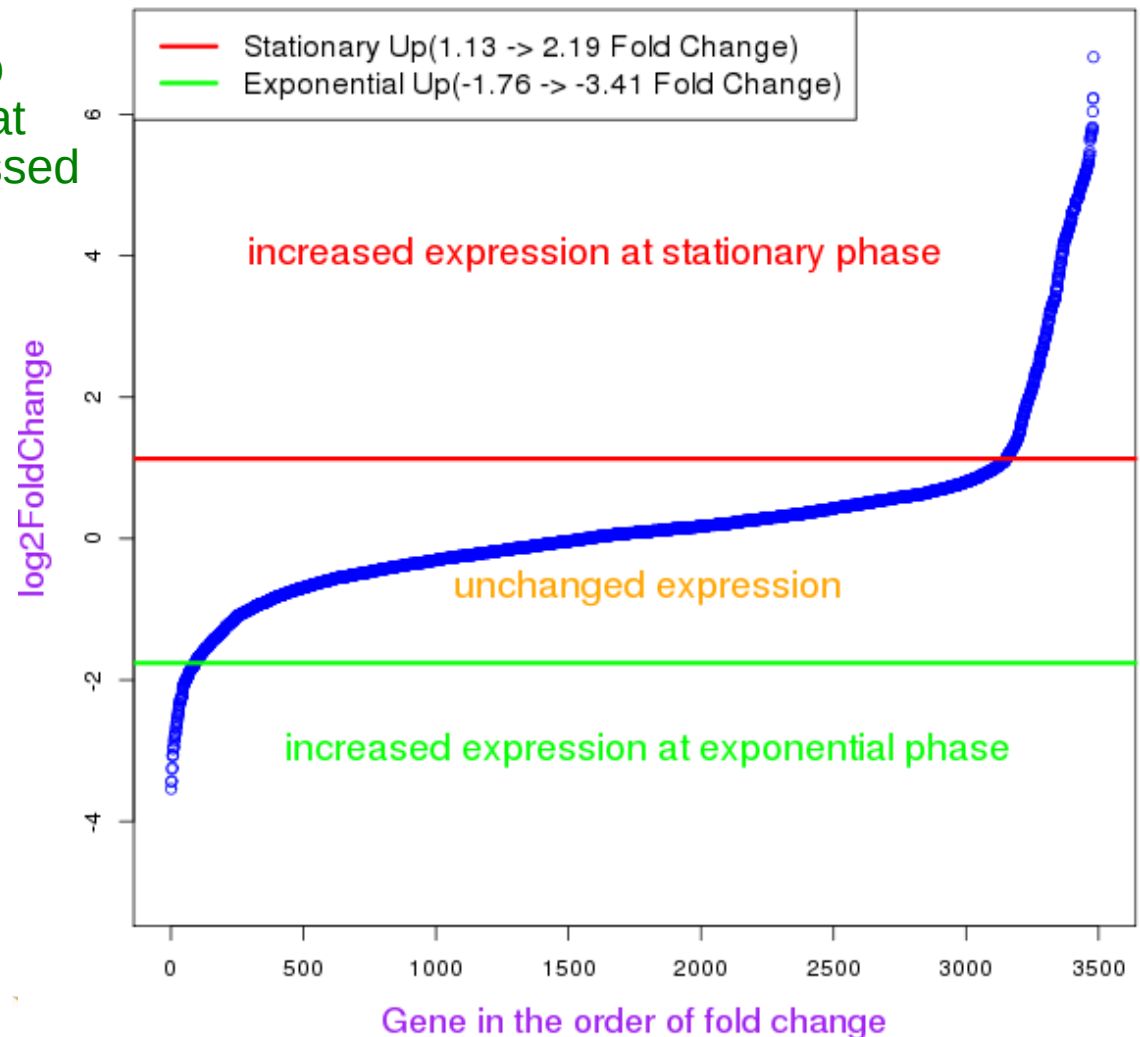
Gene expression - a moving target

How much does expression have to change, between two treatments, before we say that a gene is differentially-expressed between the two treatments?

Fold Change Test - typically uses an arbitrary cutoff of 2.0

By taking the derivative of the curve (ie. the elbow), we can choose a cutoff both for up-regulated and down-regulated genes.

Rationale of Elbow Method



Gene expression - a moving target

C. thermocellum DSM1237 - exponential vs. stationary growth

COG/Significant genes of each method

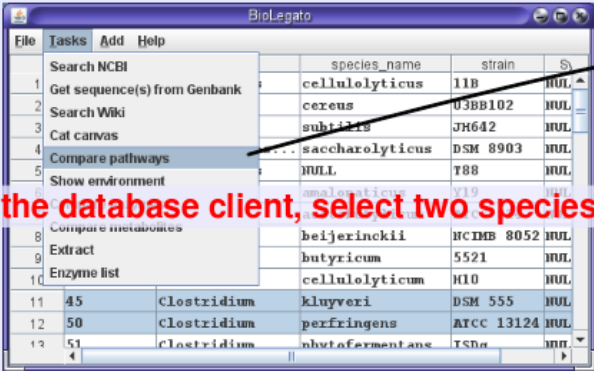
COG/Significant genes of each method	Total Genes	Elbow Method		Fold-2 Method		Cuffdiff Method	
		Up	Down	Up	Down	Up	Down
[N] Cell motility	95	5	1	53	6	29	2
[T] Signal transduction mechanisms	171	6	1	60	29	22	5
[G] Carbohydrate transport and metabolism	145	9	5	52	38	28	13
[M] Cell wall/membrane/envelope biogenesis	171	13	2	44	44	23	17
[K] Transcription	175	9	4	38	43	20	17
[S] Function unknown	188	2	1	52	53	15	15
[Z] Cytoskeleton	1	0	0	0	0	0	0
[O] Posttranslational modification, protein turnover, chaperones	89	1	2	22	23	6	8
[U] Intracellular trafficking, secretion, and vesicular transport	60	0	5	21	22	11	10
[L] Replication, recombination and repair	261	1	1	45	45	4	11
[V] Defense mechanisms	43	1	1	8	15	3	6
[Q] Secondary metabolites biosynthesis, transport and catabolism	18	0	2	3	10	0	6
[D] Cell cycle control, cell division, chromosome partitioning	38	1	1	4	16	2	6
[P] Inorganic ion transport and metabolism	93	5	1	11	28	7	12
[H] Coenzyme transport and metabolism	104	2	2	7	44	3	10
[I] Lipid transport and metabolism	46	0	7	2	29	0	20
[F] Nucleotide transport and metabolism	62	0	3	4	37	1	24
[R] General function prediction only	266	4	10	42	96	14	36
[C] Energy production and conversion	116	1	16	10	68	4	39
[E] Amino acid transport and metabolism	166	0	17	12	85	4	47
[J] Translation, ribosomal structure and biogenesis	165	2	47	11	119	3	86
total gene	2473	62	129	501	850	199	390

Gene expression - Lessons learned


- Be prepared to write your own scripts to complete your data pipeline
- Microarrays demand 6 biological replicates
- Microarrays - be prepared to throw out slides
- RNAseq largely replace microarrays
- It is a waste of time and money to do so few replicates that you have no publishable results at the end. Do fewer experiments, but do more biological replicates.

Getting lost in the systems biology pathways

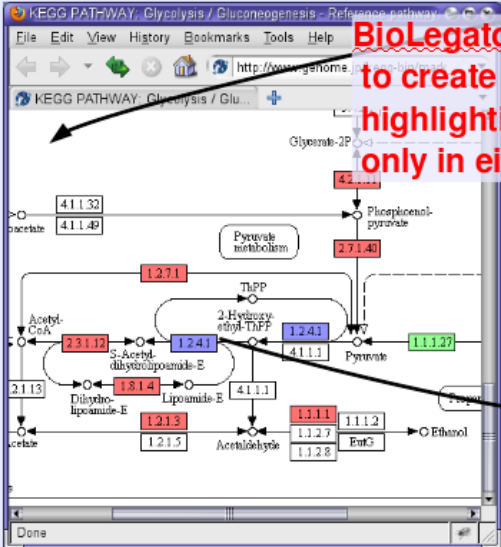
In the database client, select two species to compare



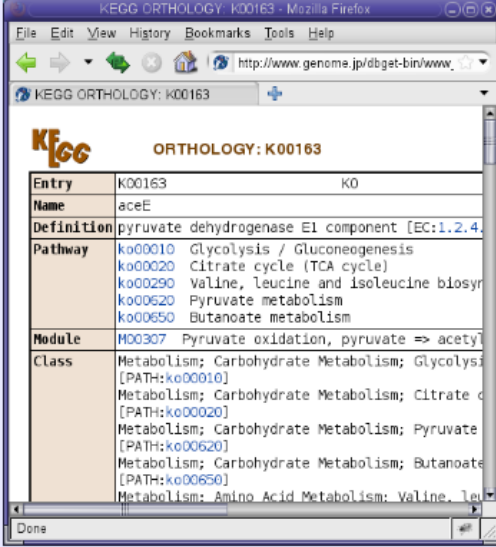
Choose pathway and click RUN



BioLegato calls the KEGG [3] API to create a custom KEGG map, highlighting enzymes present only in either species, or both

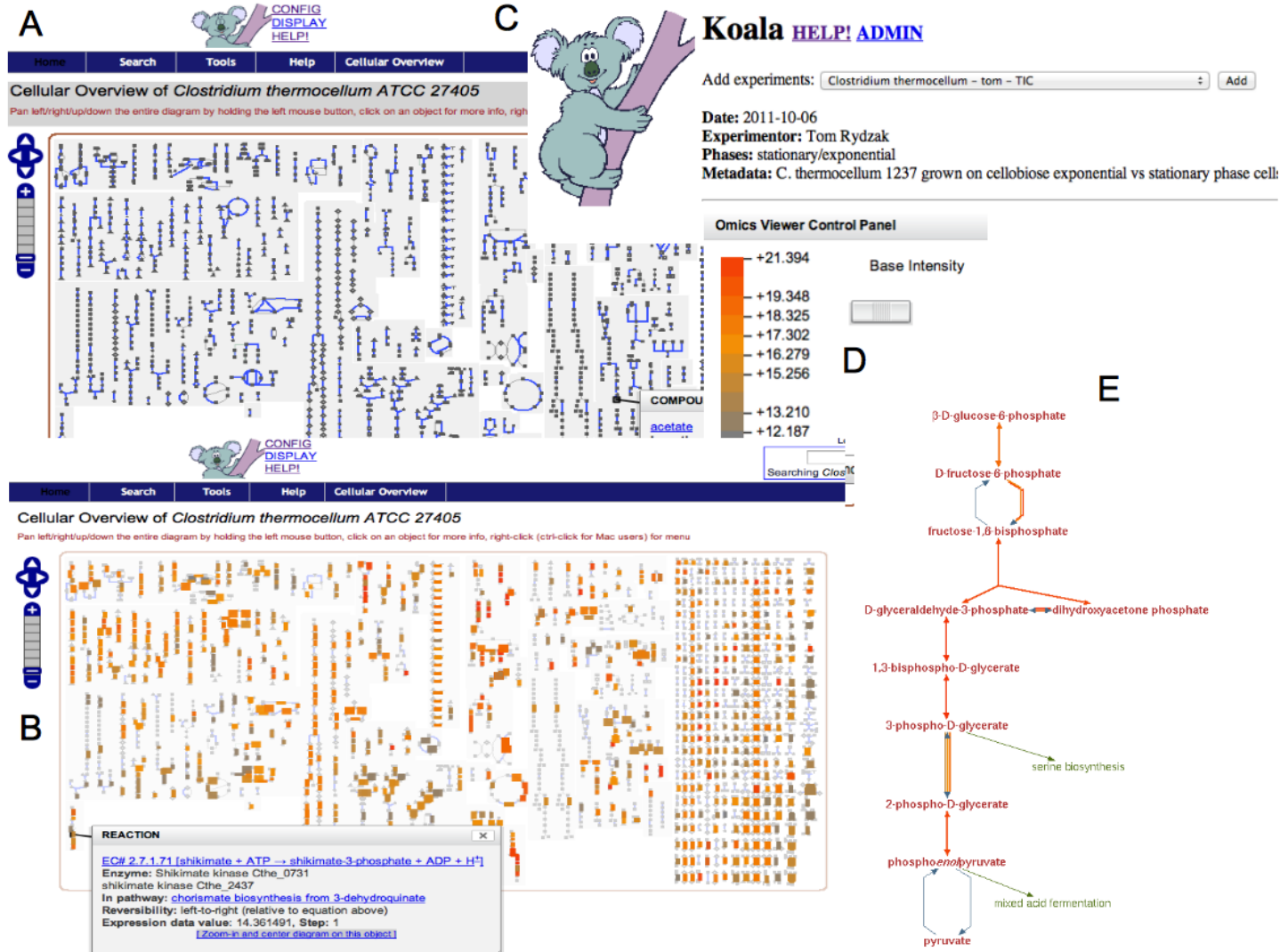


Click on an enzyme to display KEGG info



Getting lost in the systems biology pathways

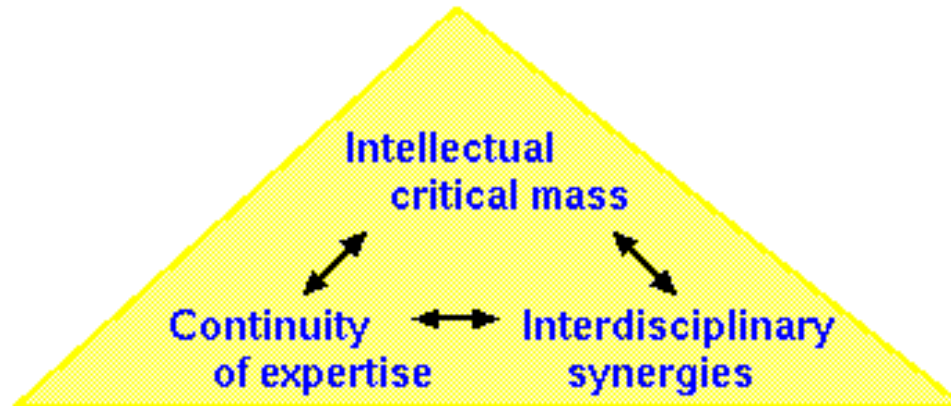
Proteomic z scores for *C. thermocellum* on cellobiose, comparing exponential vs. stationary cells



Biologists and Bioinformaticians



UNIVERSITY
OF MANITOBA



Faculty
Brian Fristensky

Bioinformaticians
Graham Alvare
Justin Zhang
Maryam Ayat

PhD Student
Abiel Roche

Research Associate
Natalie Bjorklund
Ruming Li

Undergraduate Students
Dale Hamel
Drexler Hernandez

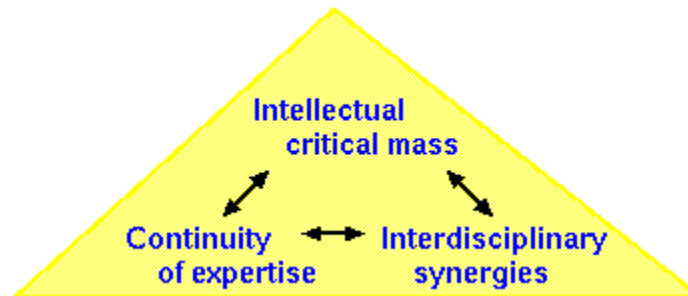
Biologists and Bioinformaticians - Lessons learned

- no one person knows everything
- There is no substitute for cross-disciplinary training.
 - Train Computer Scientists in biology
 - Train Biologists in Computer Science
- Avoid strict compartmentalization of projects. Get everybody to work together on different tasks.
- Meetings are for kicking around ideas. Minimize formal presentations
- Work with biologists on a day to day basis

The Bio Information Technologies Lab (Bit)



Bio Information Technologies
Laboratory



The Bio Information Technologies Lab (Bit)

History:

Genome Canada
Bioinformatics
Innovation Centre
(BIC) (2003 - 2011)

MGCB2 project (2008 -
2013)

Bit Lab (2012 - present)

The screenshot shows the website of the Bio Information Technologies Laboratory (Bit) in a Mozilla Firefox browser window. The browser's address bar shows the URL `home.cc.umanitoba.ca/~frist/Bit/index.html`. The website has a green header with the Bit logo and the text "Bio Information Technologies Laboratory". Below the header is a navigation bar with links: "About Us", "Services and Collaborations", "Resources", "Research and Software", "Publications and Presentations", "Training", "People", and "Contact". The main content area is divided into two columns. The left column is titled "Bioinformatics Services, Resources and Collaborations" and lists several services: "Genome assembly and annotation", "Microarray/Transcriptomics", "Systems Biology/Pathway analysis", "Databases", "Data pipelines", "Bioinformatics software", "Custom software and programming", "Project Wikis", and "Lab group computer management". Below this list is a green text box that says "Let us help you write the bioinformatics component of your next grant proposal." The right column features a table with columns for "NAME", "ACC", "PROTEIN", "SPECIES", "BUILD", "GO TERM", and "GO ID". The table contains several rows of data. Below the table is a diagram showing a yellow triangle with the text "Intellectual critical mass" at the top, and "Continuity of expertise" and "Interdisciplinary synergies" at the base. To the right of the triangle is a flowchart with numbered nodes (1-10) and arrows indicating connections. At the bottom of the page, a footer states: "BIT facilities are provided through the generosity of the Department of Computer Science and the Faculty of Science."

Bioinformatics Services, Resources and Collaborations

- Genome assembly and annotation
- Microarray/Transcriptomics
- Systems Biology/Pathway analysis
- Databases
- Data pipelines
- Bioinformatics software
- Custom software and programming
- Project Wikis
- Lab group computer management

Let us help you write the bioinformatics component of your next grant proposal.

BIT facilities are provided through the generosity of the Department of Computer Science and the Faculty of Science.

The Bio Information Technologies Lab (Bit)

Don't reinvent the wheel....



Work with the Bit lab



The Bio Information Technologies Lab (Bit)

We work with you to write the grant:

- **Benefit from our experience in writing grants for genomics projects**
- **Add credibility to your proposal**
- **Increase your chances of getting funded**

Leverage an experienced team

- **biologists**
- **bioinformaticians**
- **computer scientists**

leverage our infrastructure and resources

Get results from day 1

The Bio Information Technologies Lab (Bit)

BIRCH/BioLegato

It is MUCH harder to write a program with a graphic user interface (GUI) than to write a program that just crunches numbers.

BioLegato is a programmable graphic interface.

Create new GUIs by creating small files using the BioPCD language.

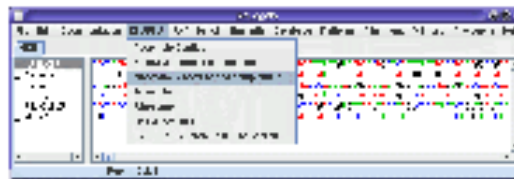
birch - application launcher



birchadmin - BIRCH administration tool



bidna - DNA sequences



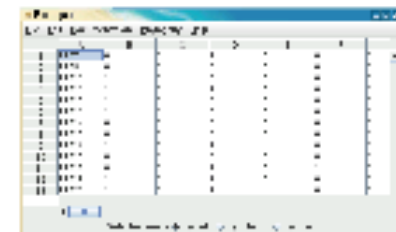
bitree - phylogenetic trees



biprotein - protein sequences

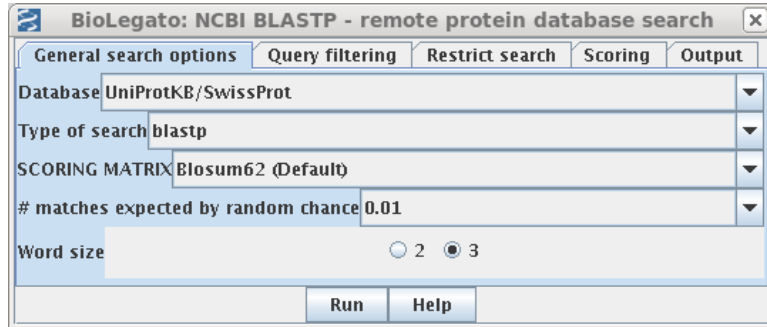


bimarker - molecular markers



The Bio Information Technologies Lab (Bit)

Menu to set parameters and run program:



Code:

tabset

```
# - - - - -
tab "General search options"
var "dbase"
    type      combobox
    label     "Database"
    default   0
    choices   "UniProtKB/SwissProt" "swissprot"
              "GenBank NonRedundant (Protein)" "nr"
              "Reference proteins" "refseq_protein"
              "Protein Structure Data Bank (PDB)" "pdb"
              "GenBank Patented" "pat"
              "Metagenomic proteins" "env_nr"
```

*International Journal of Computer Applications (0975 – 8887)
Volume 57– No.6, November 2012*

BioPCD - A Language for GUI Development Requiring a Minimal Skill Set

Graham GM Alvare
Dept of Plant Science
University of Manitoba
Winnipeg, MB. Canada. R3T 2N2

Abiel Roche-Lima
Dept of Computer Science
University of Manitoba
MB. Canada. R3T 2N2

Brian Fristensky
Dept of Plant Science
University of Manitoba
Winnipeg, MB. Canada. R3T 2N2

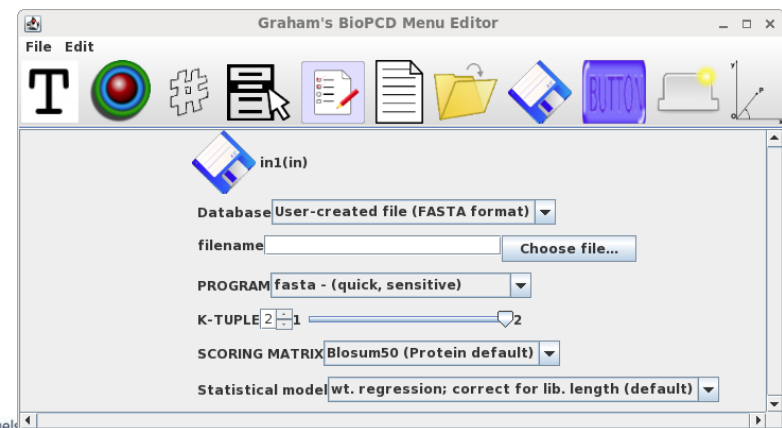
ABSTRACT

BioPCD is a new language whose purpose is to simplify the creation of Graphical User Interfaces (GUIs) by biologists with minimal programming skills. The first step in developing BioPCD was to create a minimal superset of the language referred to as PCD (Pythonesque Command Description). PCD defines the core of terminals and high-level non-terminals required to describe data of almost any type. BioPCD adds to PCD the constructs necessary to describe GUI components and the syntax for executing system commands. BioPCD is implemented using JavaCC to convert the grammar into code. BioPCD is designed to be terse and readable and simple enough to be learned by copying and modifying existing BioPCD files. We demonstrate that BioPCD can easily be used to generate GUIs for existing command line programs. Although BioPCD was designed to make it easier to run bioinformatics programs, it could be

many languages. Because of rapid advancement in the field, there are always new programs appearing in the literature that replace the current existing "favorite" of researchers working in that field. The work required to modify an existing GUI to handle the new program is also a limiting factor. An additional problem is that many programs are not written with extensibility in mind.

The need for easy ways to create GUIs is particularly important in bioinformatics. Bioinformatics is an interdisciplinary field, in which computer methodologies are applied to the analysis of biological data. The most critical limiting factor in bioinformatics is not computational, but rather human. The complexity of the data, along with the enormous datasets generated, often push the limits of computer resources and algorithmic rigor. However, few biologists have any formal training in computers. Thus, the user group with the greatest need for simple GUI interfaces to

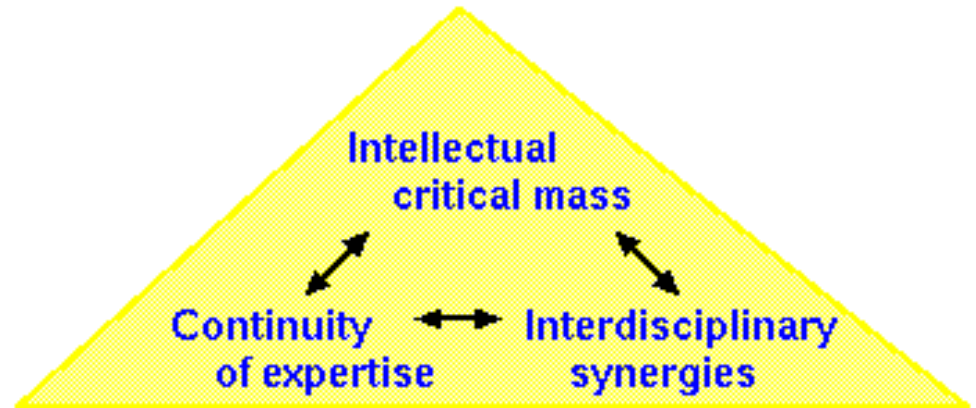
Point-and-click PCD editor



The Bio Information Technologies Lab (Bit)

Industry level standards for

- data security
- data integrity and backups
- server availability
- software engineering



We're bilingual. We speak:

- the language of biology
- the language of bioinformatics

Research Partner Logos



UNIVERSITY
OF MANITOBA



GenomeCanada



GenomePrairie



Canada Foundation
for Innovation
Fondation canadienne
pour l'innovation



Agriculture and
Agri-Food Canada

Agriculture et
Agroalimentaire Canada



National Renewable
Energy Laboratory

UNIVERSITY OF
Waterloo



Government of
Saskatchewan



UNIVERSITY OF
SASKATCHEWAN



iidd

Institut
international du
développement
durable

International
Institute for
Sustainable
Development



Ontario Genomics Institute