Genome Assembly and Other Analyses using Next Generation Sequence Data

Robert Bruccoleri
Congenomics, LLC

Outline

- Introduction and Prerequisites
- Practical examples of NGS applications in biotechnology and pharmaceuticals:
  - Denovo Assembly of Bacterial Genomes
  - Escapades in Plant Genome Assembly
  - Tracking Selective Evolution of Antibody Domains
- Summary of Advice

Congenomics, LLC

- Congenomics, LLC is a Bioinformatics and Genomics consulting company.
- Formal training in Computer Science, Molecular Biology, and Computational Chemisty.
- Experience in Molecular Modeling, Scientific Computing, Bioinformatics, and Genomics.

Prerequisites

- Biology and Computer Science are required equally.
  - Understand the biological science
  - Program and administer the computing resources
- Appreciation for the strengths and weaknesses of technology and protocols is also vital.
- Enjoy being on the frontiers of biological science.

Bacterial genome assembly using next generation sequencing

Robert Bruccoleri, Joseph Szustakowski, Edward Oakeley, Michael Derby, Charlie Moore

NOVARTIS

Outline

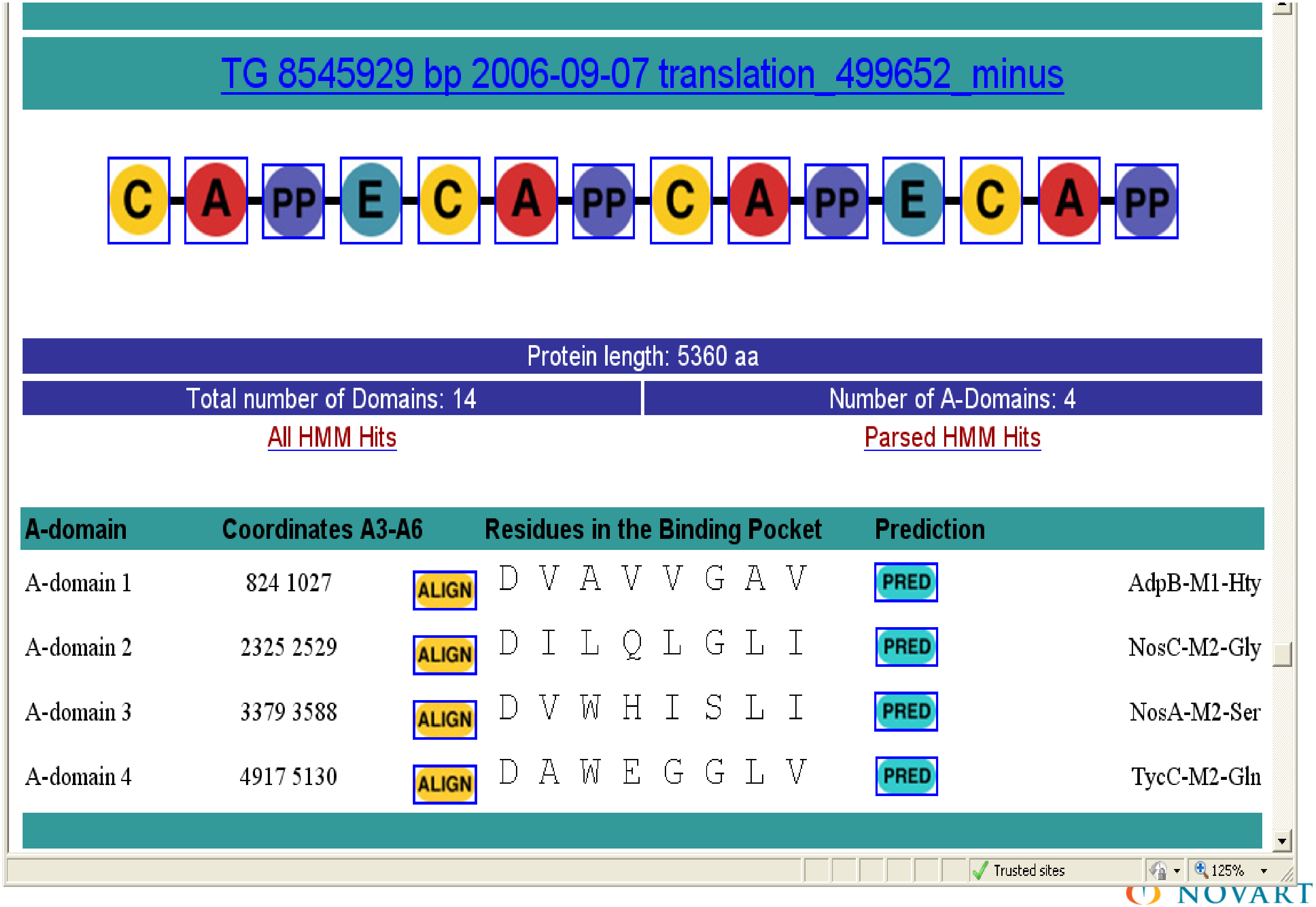
- Project Background and goals
- *De novo* assembly bioinformatics
- Details of a genome assembly
- Summary and lessons learned
- Future plans

Genome Sequencing for Natural Product Research

Different classes of microorganisms like bacteria of the classes actinomycetes and myxobacteria as well as filamentous fungi are important sources for the discovery of novel natural products.

1. Identification of Non-ribosomal Peptide Synthases (NRPS) and PolyKetide Synthases (PKS).
2. Structure prediction based on biosynthetic clusters.
3. Match biosynthetic clusters to known compounds.
4. Silent NRPS and PKS gene cluster evaluation.

Non-ribosomal Peptide Synthases are Very Large Multi-Domain Proteins.



TG 8545929 bp 2006-09-07 translation_499652_minus

Protein length: 5360 aa

Total number of Domains: 14

Number of A-Domains: 4

All HMM Hits

Parsed HMM Hits

A-domain	Coordinates A3-A6		Residues in the Binding Pocket	Prediction	
A-domain 1	824 1027	ALIGN	D V A V V G A V	PRED	AdpB-M1-Hty
A-domain 2	2325 2529	ALIGN	D I L Q L G L I	PRED	NosC-M2-Gly
A-domain 3	3379 3588	ALIGN	D V W H I S L I	PRED	NosA-M2-Ser
A-domain 4	4917 5130	ALIGN	D A W E G G L V	PRED	TycC-M2-Gln

Project Goals

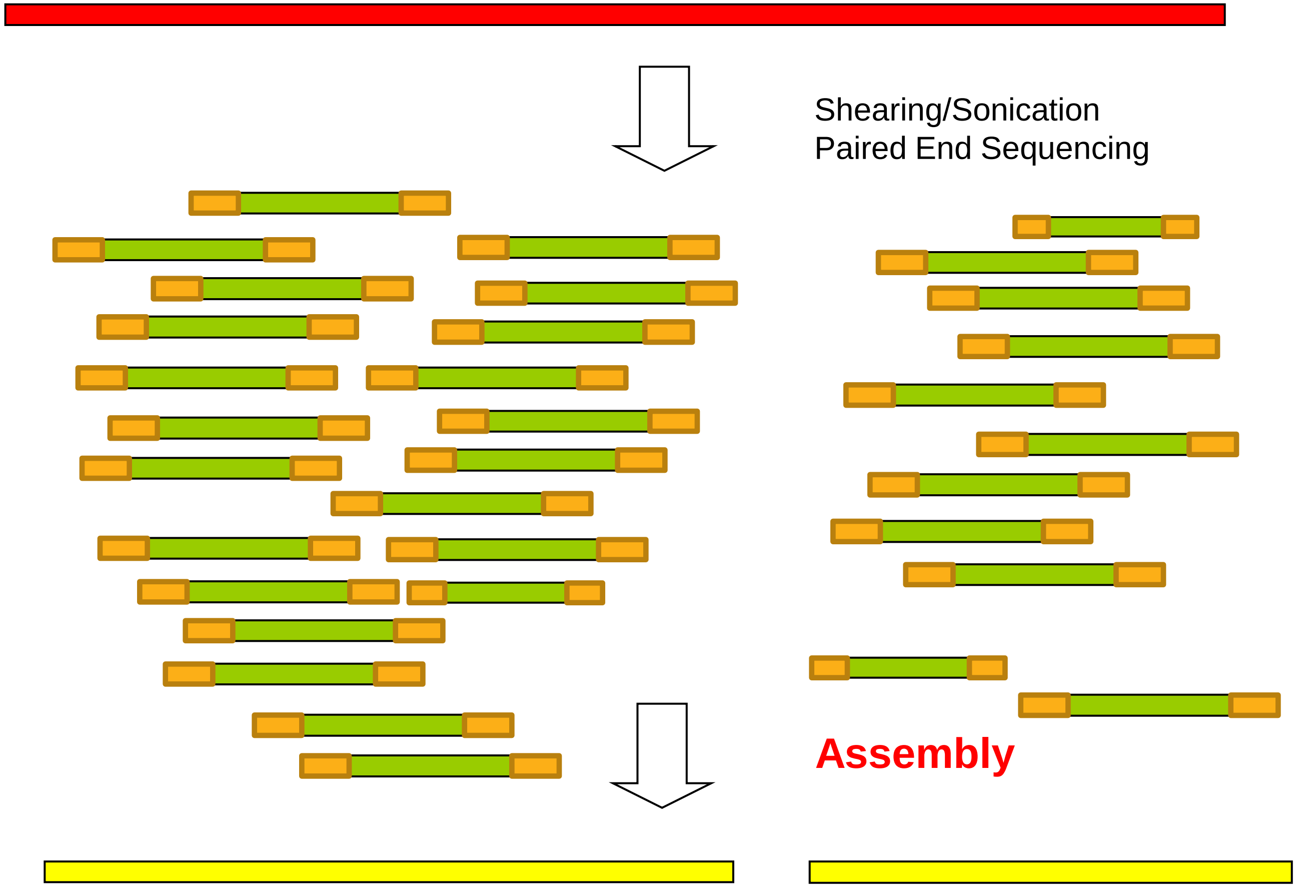
- Sequence *bacterium #1*, evaluating various NGS platforms in the process.

- Perform a *de novo* assembly of *bacterium #1*.

- Identify the NRPS gene cluster that synthesizes cyclic peptide #1.

- Big picture: Can we establish a sequencing and bioinformatics platform that would support future *de novo* microbial sequencing projects?

Computational Problem



Genomic DNA
10 Mbp

Shearing/Sonication
Paired End Sequencing

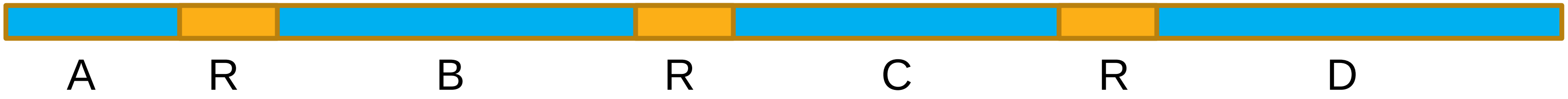
Shotgun reads:
100 M pairs
~500bp each

Assembly

Contigs

From Oswaldo Cruz Institute, Department of Biochemistry and Molecular Biology

Repetitive DNA



Reads: AR RB BR RC CR RD

Possible contigs

ARBRCRD
ARCRBRD
ARBRBRBRBRCRD
ARD
RBRD

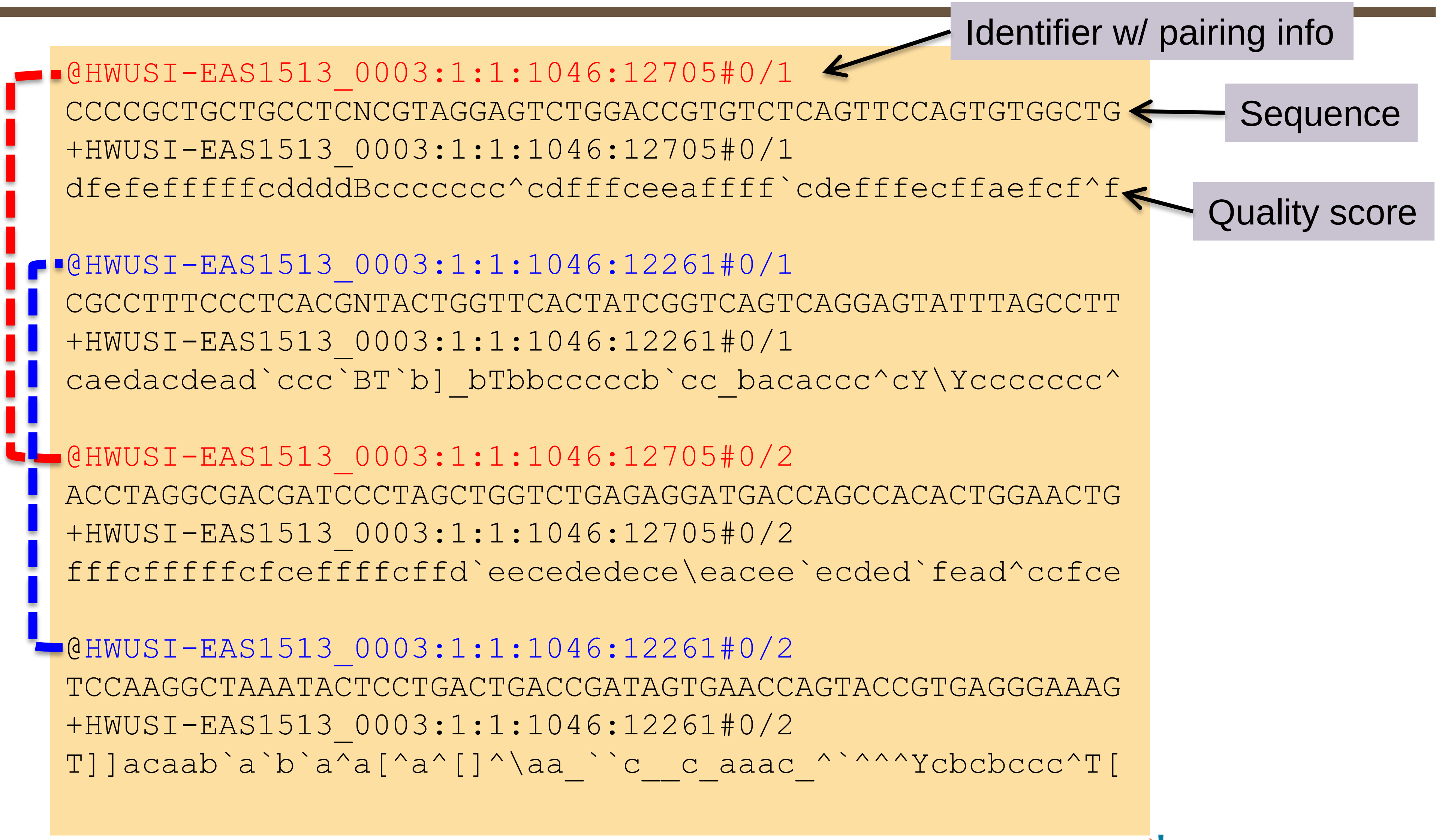
Partial resolution: Assume uniform read distribution and treat higher coverage regions as repeats. Repeats are not extended.

Best solution: Get longer reads and read pairs on larger fragments.

Sequence sources

- Illumina: 36-100 bp reads. Paired ends possible with separation up to 1000 bp. Inexpensive. In house.
- SOLiD: 50 bp reads of dinucleotides (color space) with highly variable quality. Least expensive. In house.
- 454/Roche: Pyrosequencing reads over 400bp. Homopolymers problematic. Most expensive. Outside service provider.
- All Next Generation sequencing technology is full of errors – this is messy work. Redundancy corrects the errors.

Example Illumina Data



```
@HWUSI-EAS1513_0003:1:1:1046:12705#0/1
CCCCGCTGCTGCCTCNCGTAGGAGTCTGGACCGTGTCTCAGTTCCAGTGTGGCTG
+HWUSI-EAS1513_0003:1:1:1046:12705#0/1
dfefefffffcddddBccccccc^cdfffceeaffff`cdefffecffaefcf^f

@HWUSI-EAS1513_0003:1:1:1046:12261#0/1
CGCCTTTCCCTCACGNTACTGGTTCACTATCGGTCAGTCAGGAGTATTTAGCCTT
+HWUSI-EAS1513_0003:1:1:1046:12261#0/1
caedacdead`ccc`BT`b]_bTbbcccccb`cc_bacaccc^cY\Yccccccc^

@HWUSI-EAS1513_0003:1:1:1046:12705#0/2
ACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTG
+HWUSI-EAS1513_0003:1:1:1046:12705#0/2
fffcfffffcfceffffcffd`eecededece\eacee`ecded`fead^ccfce

@HWUSI-EAS1513_0003:1:1:1046:12261#0/2
TCCAAGGCTAAATACTCCTGACTGACCGATAGTGAACCAGTACCGTGAGGGAAAG
+HWUSI-EAS1513_0003:1:1:1046:12261#0/2
T]]acaab`a`b`a^a[^a^[]^\aa_``c__c_aaac_^`^^^Ycbcbccc^T[
```

Identifier w/ pairing info

Sequence

Quality score

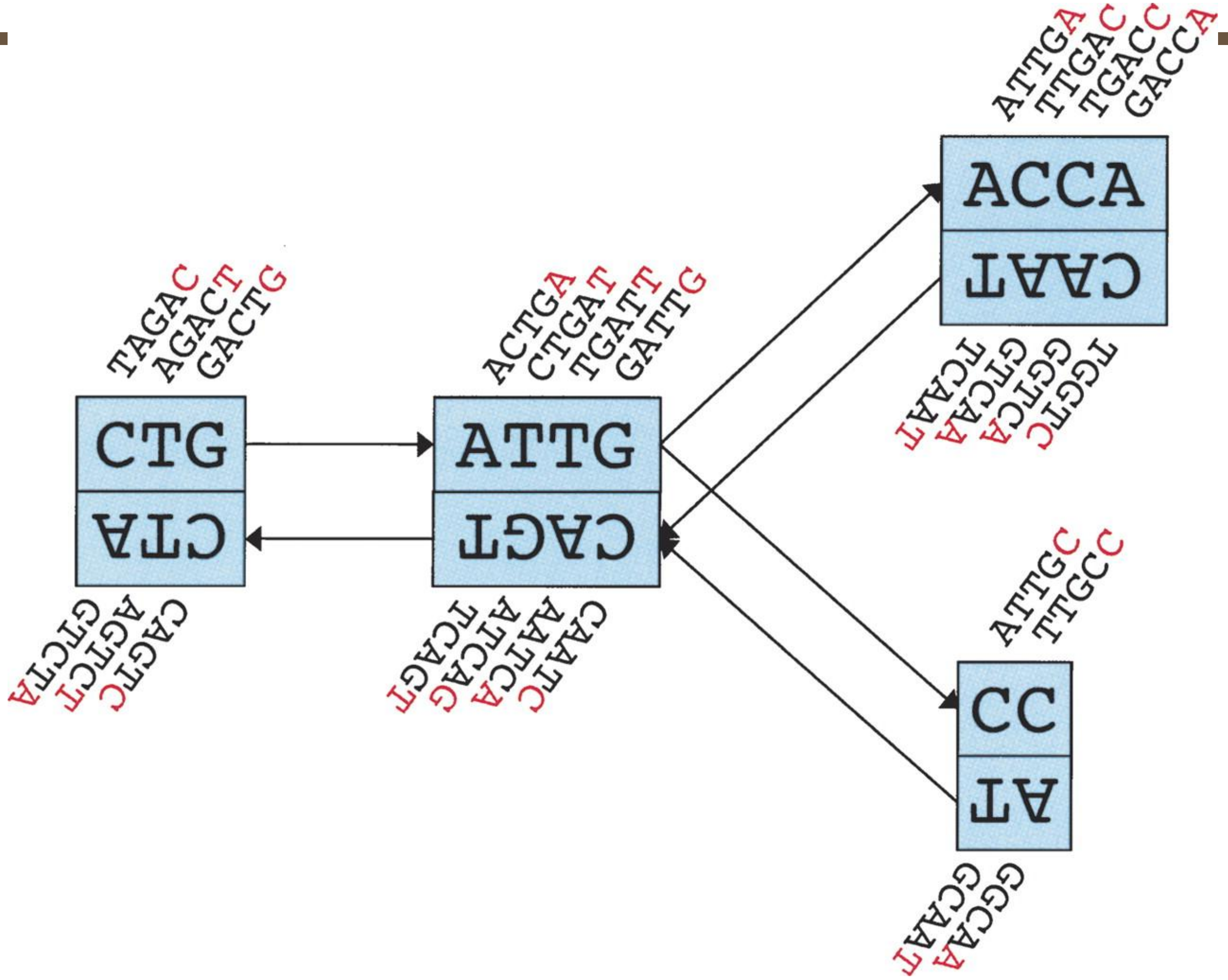
Information at Project Initiation

- Bacterium #1 has large genome (about 10 Mbp) extremely GC rich.
- 192,301,536 reads, 55bp, all paired end with inserts between 400 to 600 bp.
- No close relatives sequenced.
- Amino acid sequence for cyclic peptide #1 was known.

Available Algorithms and Software

- de Bruijn graphs:
  - Velvet (GPL): Contigs were short, and quality scores were not used.
  - ABySS (Non-commercial only): $10K/year/site, time consuming licensing process.
  - CLC Bio (Commercial): Very fast, but inconsistent. No record of process.
- Overlap Graphs
  - MIRA (GPL): Slow, memory intensive, but effective. 14 years of development. Copious record of process. Author works for a company that depends on sequencing for its business. Well supported.

Schematic representation of our implementation of the de Bruijn graph.

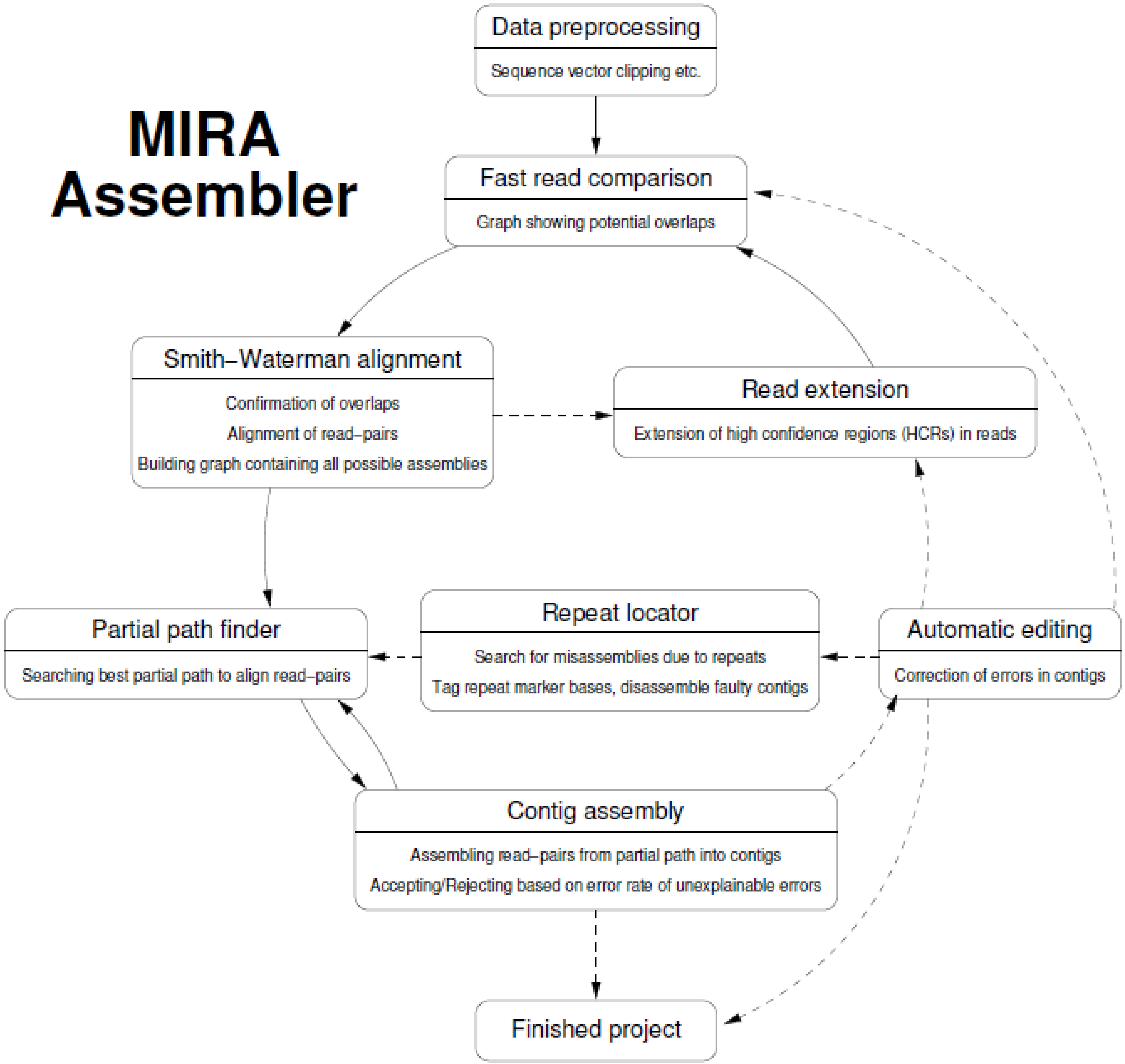


Zerbino D R , Birney E Genome Res. 2008;18:821-829

Available Algorithms and Software

- de Bruijn graphs:
  - Velvet (GPL): Contigs were short, and quality scores were not used.
  - ABySS (Non-commercial only): $10K/year/site, time consuming licensing process.
  - CLC Bio (Commercial): Very fast, but inconsistent. No record of process.
- Overlap Graphs
  - MIRA (GPL): Slow, memory intensive, but effective. 14 years of development. Copious record of process. Author works for a company that depends on sequencing for its business. Well supported.

MIRA Assembler



Data preprocessing

Sequence vector clipping etc.

Fast read comparison

Graph showing potential overlaps

Smith–Waterman alignment

Confirmation of overlaps

Alignment of read–pairs

Building graph containing all possible assemblies

Read extension

Extension of high confidence regions (HCRs) in reads

Partial path finder

Searching best partial path to align read–pairs

Repeat locator

Search for misassemblies due to repeats

Tag repeat marker bases, disassemble faulty contigs

Automatic editing

Correction of errors in contigs

Contig assembly

Assembling read–pairs from partial path into contigs

Accepting/Rejecting based on error rate of unexplainable errors

Finished project

Initial Attempts(1)

- Large number of reads was too much for Mira or Velvet, selection of subsets was necessary.
- Simultaneous assembly of all subsets of 7 million reads crashed multiple servers on our computer cluster.
  - NIBR IT high performance computing group provided a dedicated queue on a high memory (96 GB RAM) server to enable this work.
  - 10 terabytes of intermediate storage used for assemblies.
  - Runs generate large amounts of intermediate files, log files.
  - We are much more diligent about removing intermediate files now.

Initial Attempts(2)

- Sorting by quality scores and using the highest quality reads did not work:
  - Because high GC content produces lower quality reads, highest quality reads had lower than average GC content and were not representative of the bacterial genome.
- All assembly algorithms generated contigs < 50kbp.
- Contigs only contained fragments of the expected NRPS
- Assemblies take ~ 1 day.

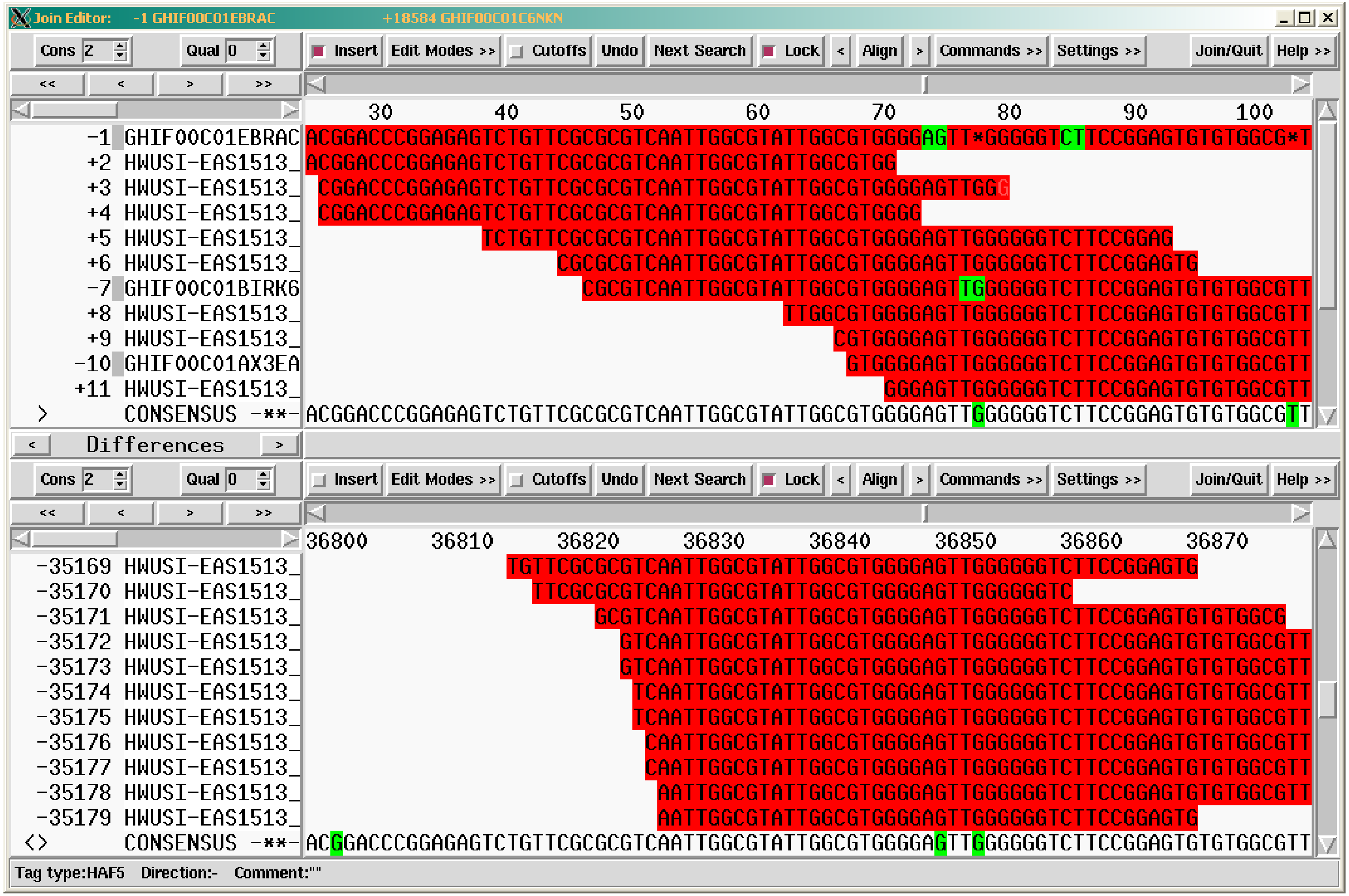
Breakthrough

- On June 1, we obtained 624,330 Roche/454 reads using Titanium reagents (approx. 500bp reads).
- Hybrid assemblies using MIRA on small fractions of the Illumina reads (7 million) with the full set of 454 reads were much better.
  - Largest contig > 500,000 bp
  - N50 > 100,000 bp
- This assembly should have been good enough to spot the NRPS.

Finishing the gene

- NRPS analysis (by PKS/NRPS program of Jacques Ravel) revealed two contigs that had protein sequences predicted to be adenylation domains with binding specificities matching disjoint parts of the cyclic peptide #1.
- Blast of these contigs found 3’ of one matched 5’ end of another, and an internal repeat in one of the contigs.
- The two contigs were joined by hand and the resulting contig produced an NRPS whose predicted substrate sequence was clearly the best match to cyclic peptide #1.

Assembly illustration



Join Editor: -1 GHIF00C01EBRAC +18584 GHIF00C01C6NKN

```
                            30        40        50        60        70        80        90      100
    -1 GHIF00C01EBRAC ACGGACCCGGAGAGTCTGTTCGCGCGTCAATTGGCGTATTGGCGTGGGGAGTT*GGGGGTCTTCCGGAGTGTGTGGCG*T
    +2 HWUSI-EAS1513_ ACGGACCCGGAGAGTCTGTTCGCGCGTCAATTGGCGTATTGGCGTGG
    +3 HWUSI-EAS1513_  CGGACCCGGAGAGTCTGTTCGCGCGTCAATTGGCGTATTGGCGTGGGGAGTTGGG
    +4 HWUSI-EAS1513_  CGGACCCGGAGAGTCTGTTCGCGCGTCAATTGGCGTATTGGCGTGGGG
    +5 HWUSI-EAS1513_               TCTGTTCGCGCGTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAG
    +6 HWUSI-EAS1513_                     CGCGCGTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTG
    -7 GHIF00C01BIRK6                       CGCGTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
    +8 HWUSI-EAS1513_                                       TTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
    +9 HWUSI-EAS1513_                                           CGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
   -10 GHIF00C01AX3EA                                            GTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
   +11 HWUSI-EAS1513_                                               GGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
>      CONSENSUS -**- ACGGACCCGGAGAGTCTGTTCGCGCGTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
```

```
                      36800     36810     36820     36830     36840     36850     36860     36870
-35169 HWUSI-EAS1513_                 TGTTCGCGCGTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTG
-35170 HWUSI-EAS1513_                   TTCGCGCGTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTC
-35171 HWUSI-EAS1513_                        GCGTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCG
-35172 HWUSI-EAS1513_                          GTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
-35173 HWUSI-EAS1513_                          GTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
-35174 HWUSI-EAS1513_                           TCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
-35175 HWUSI-EAS1513_                           TCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
-35176 HWUSI-EAS1513_                            CAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
-35177 HWUSI-EAS1513_                            CAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
-35178 HWUSI-EAS1513_                             AATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
-35179 HWUSI-EAS1513_                             AATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTG
<>     CONSENSUS -**- ACGGACCCGGAGAGTCTGTTCGCGCGTCAATTGGCGTATTGGCGTGGGGAGTTGGGGGGTCTTCCGGAGTGTGTGGCGTT
```

Tag type:HAF5 Direction:- Comment:""

Summary

- We have successfully sequenced and assembled bacterium #1 using NGS and bioinformatics.
- We have identified the NRPS that synthesizes peptide #1.
- We have nearly completed the sequencing and assembly of Bacterium #2 using similar techniques.
- *De novo* assembly of microbial genomes is tractable using technologies available today.

Lessons learned

- Combination of technologies was needed
  - 454 gives long reads, but has homopolymer errors.
  - Illumina provides depth of coverage and corrects homopolymer errors.
  - This assembly would not have been possible without the 454 data.
  - GC bias in PCR appears to be an issue.
- Short reads are not enough for de novo assembly. Paired ends of long fragments or long reads are needed.
- Simulations are predictive
  - Simulations using *S. griseus* yielded similar contig lengths and distributions as real data.
  - We are using simulations to guide the design of future experiments.
- Domain knowledge is essential
  - Without specific gene knowledge, we would not have focused on a specific contig and found the connection between NRPS and compound #1.
- **Close collaboration, open communication is critical.**

Future Plans(1)

- Test new methods and technologies in sequencing:
  - PCR free library preparation.
  - New Illumina chemistry for reducing GC bias on the flow cell.
  - Longer reads on longer paired ends using Illumina.
  - Combinations of methods.
  - Determine minimum requirements to reduce cost.
  - Pacific Biosciences sequencing: very long, strobed reads.
- Process improvements: introduce automation whenever possible.
- Data integration: Adapt and write software to provide information in a more useful form for our biology collaborators.

Future plans (2)

- Genome databases and annotation:
  - Develop sequence databases to capture annotation of individual genomes and microbiological knowledge about the bacteria.
- Improvements to substrate and functional predictions of NRPS and PKS domains
  - Over time, accumulate substantial information about domain specificities.
  - Develop and/or improve functional predictions like Jacques Ravel's NRPS program.
  - Possibly partner with academic groups to improve predictions.

Software Resources

- MIRA: http://sourceforge.net/projects/mira-assembler/
- NRPS/PKS: http://sourceforge.net/projects/secmetdb/
- GAP4: http://sourceforge.net/projects/staden/ Displays assembly data.
- CLVIEW: http://compbio.dfci.harvard.edu/tgi/software/ Displays assembly information in ACE files. Easier to use and faster than GAP4, but very limited capabilities.
- MAUVE: http://gel.ahabs.wisc.edu/mauve/download.php Compares assemblies graphically.

Acknowledgments

- **BMD / Genome Technologies**

  Virginie Petitjean, Christian Kohler, Stine Büchmann-Møller, Anita Fernandez, Moritz Frei, Martin Letzkus, Frank Staedtler, Sandrine Starck-Schwertz, Sandrine Bongiovanni

- **BMD**

  Joanne Meyer, N.R. Nirmala, Keith J Johnson

- **Natural Products Unit**

  Charles M. Moore, Esther K. Schmitt, Kathrin Buntin, Klaus Memmert, Philipp Krastel, Hansueli Naegeli, Ying Wang, Tim Schuhmann, Eric Weber, and Frank Petersen

- **NIBR IT**

  Wolfgang Zipfel, Steve Litster

- **DNAVision**
- **Bastien Chevreux** (author of MIRA) - DSM Nutritional Products Switzerland
- **Jacques Ravel** (author of NRPS program) – University of Maryland School of Medicine

Escapades in Plant Genome Assembly

- A client of Congenomics LLC is interested in a handful of genes from a plant genome.
- 240 million 100bp paired end reads using Illumina GA IIx were obtained.
- Homologues are available from Arabidopsis thaliana
- Find the genes.
- Use relational database and SQL!

Iterative assembly

```
R = reads from gene of interest
For i=1 to Ncycles do
    W = words(R);
    NR1 = Reads containing W
    NR2 = NR1 union mate_pairs(NR1)
    R = R union NR2
    Assemble R using Mira
Done
```

Iterative Assembly Implementation(1)

- Use a compact integer representation of DNA alphabet: A = 0, C = 1, G = 2, T = 3.
  - Each base encoded in two bits.
  - Ambiguities and unknowns ignored.
  - 32 bp fits in 64 bit integer.
-
```
CREATE TABLE words (
          numid integer,
          twobit bigint);
```
- Index both columns

Iterative Assembly Implementation(2)

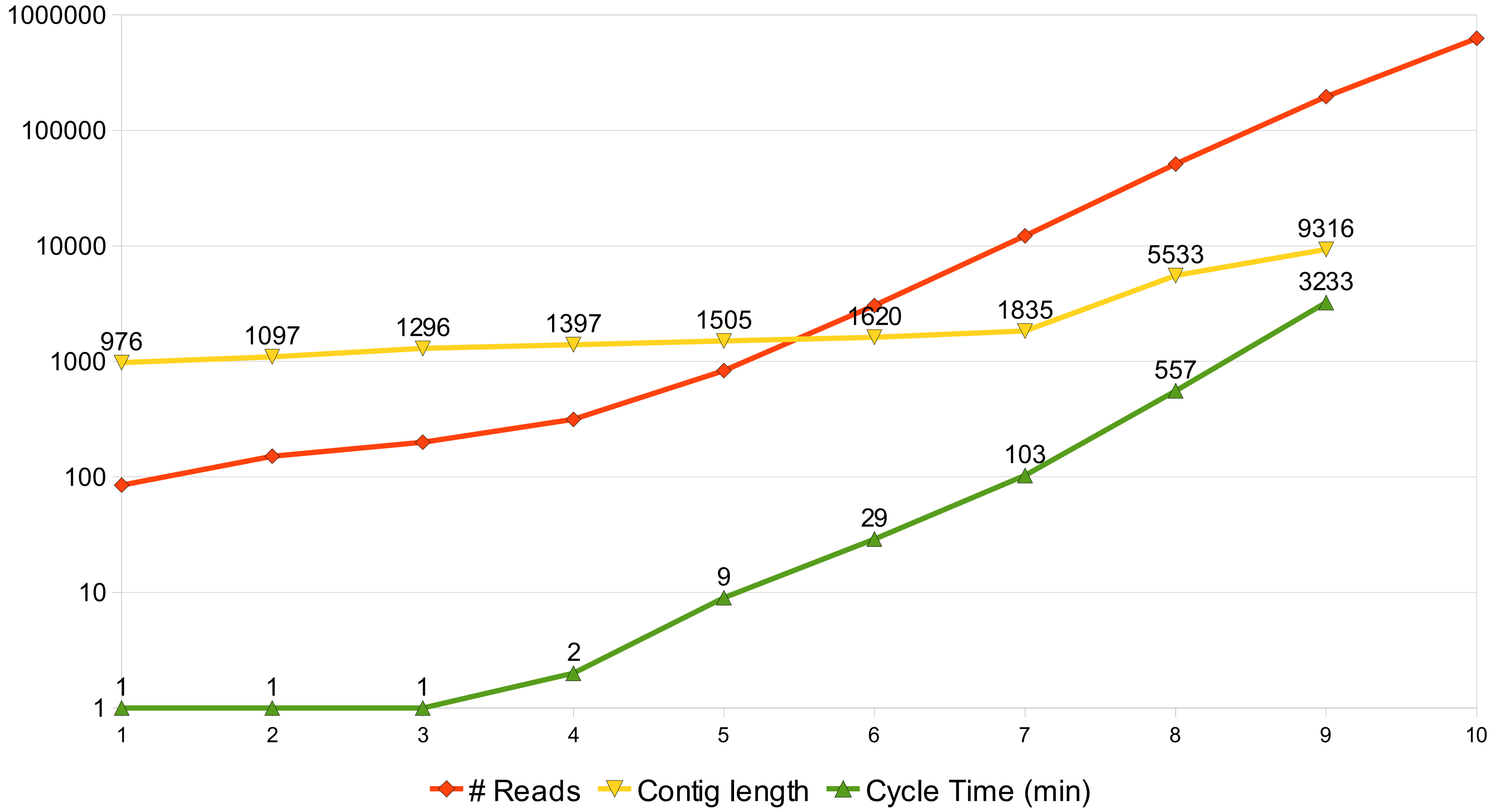
```
CREATE TABLE reads (
   numid integer,
   run text,
   id text,
   pair_code character(1),
   seq text,
   qual text);
```

Index on numid and id.

Iterative Assembly Implementation(3)

- Must also keep track of word counts:
  - Can choose to avoid assembling through repetitive sequence.
- Current database size is 904 GB

Iterative Assembly Example



Does it work?

- Partial success:
  - Genomic sequence has been extended to find some of the exons in the genes of interest
  - But, the most distant exon has not been.
- Recently, 3kbp and 5kpb mate pair data has been added to the database.
- Still a work in progress...

Lessons Learned

- Large genome assembly not possible with short read data. Physical mapping still needed.
- Gene oriented iterative assemblies *may* work – jury is out.
- Relational database storage of NGS data permits selective algorithms and is feasible.

Resources

- PostgreSQL (http://www.postgresql.org/)
- Mira (http://sourceforge.net/projects/mira-assembler)
- Frescobi (watch http://search.cpan.org) - will be released when tidied up.

NGS at X-Body Biosciences

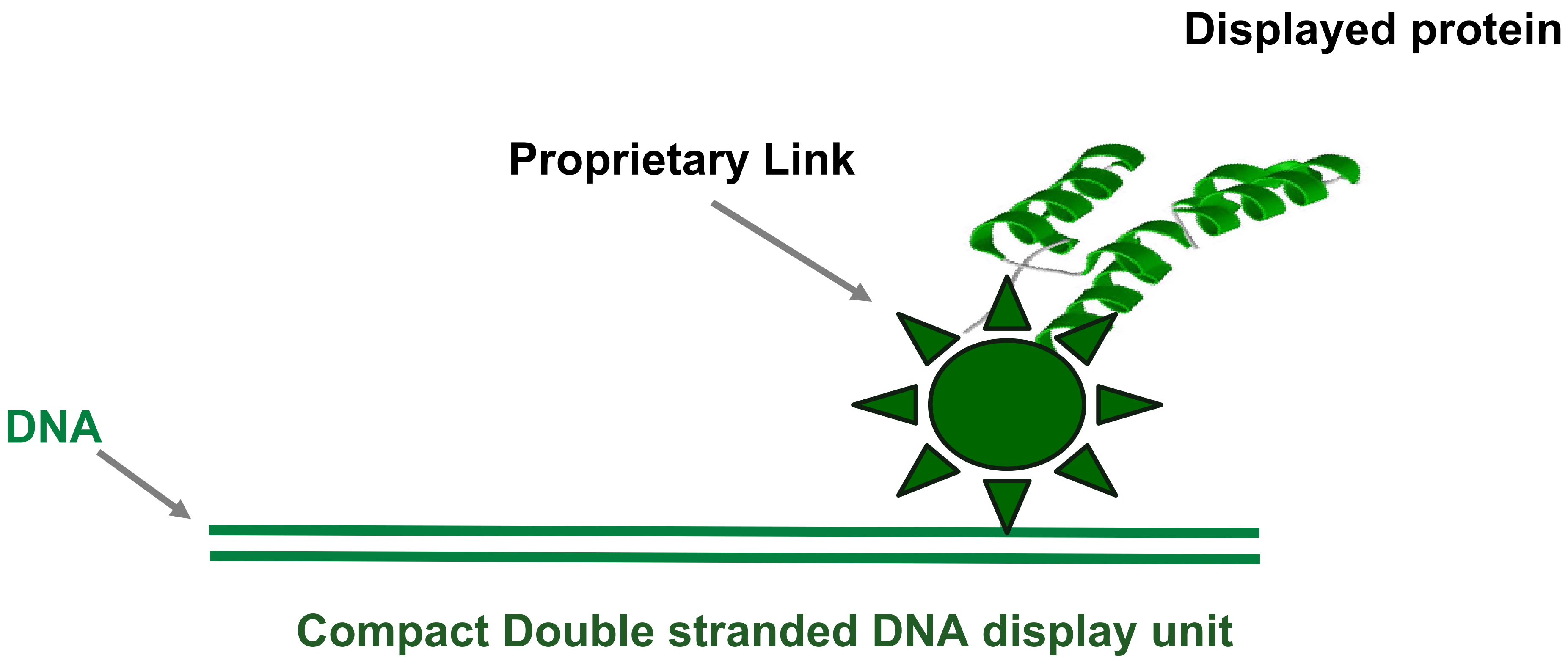
Tracking Selective Evolution of Antibody Domains

Robert Bruccoleri
Alexander Litovchick
Yan Chen
Richard Wagner
Tod Woolf

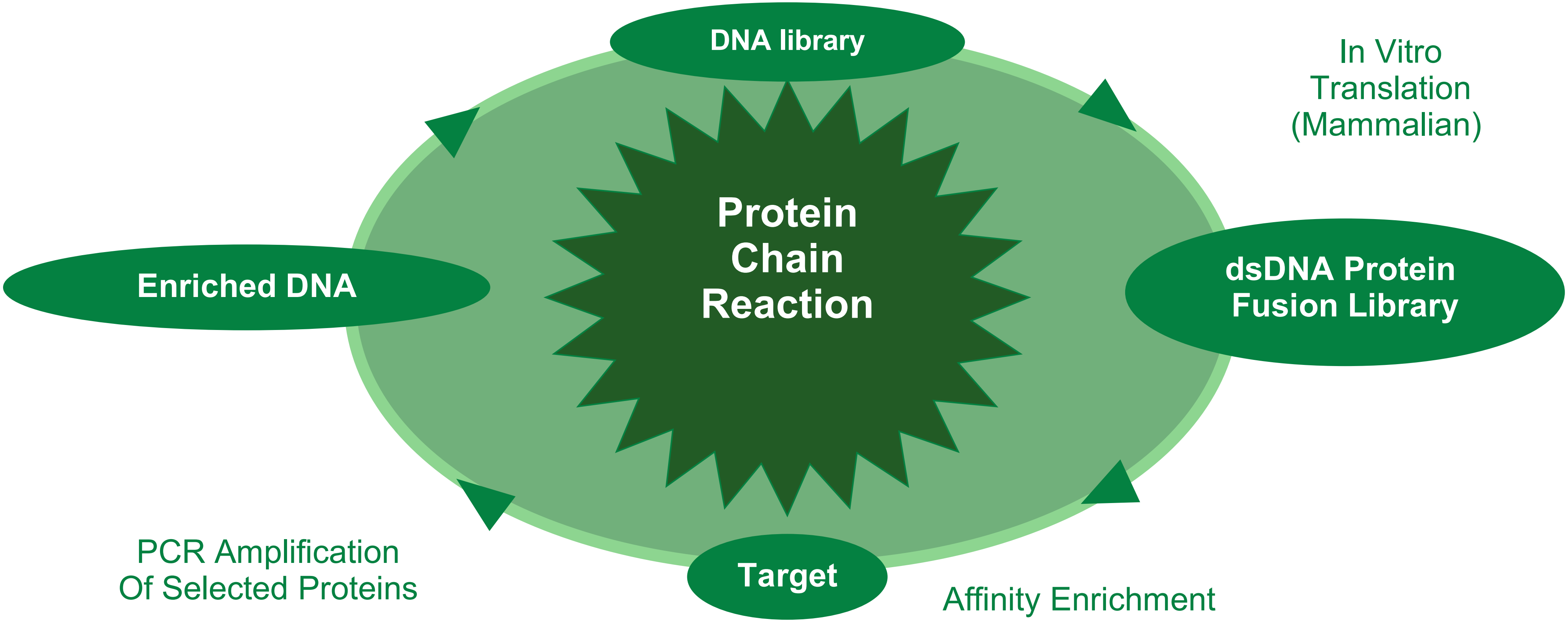
NGS at X-Body Biosciences

- Antibody domain production system and the importance of deep sequencing.
- Sequence processing
- Presentation of analysis
- Lessons learned
- Resources

X-BODY dsDNA Display Offers an Improved Alternative Path to Proven *In Vitro* Display



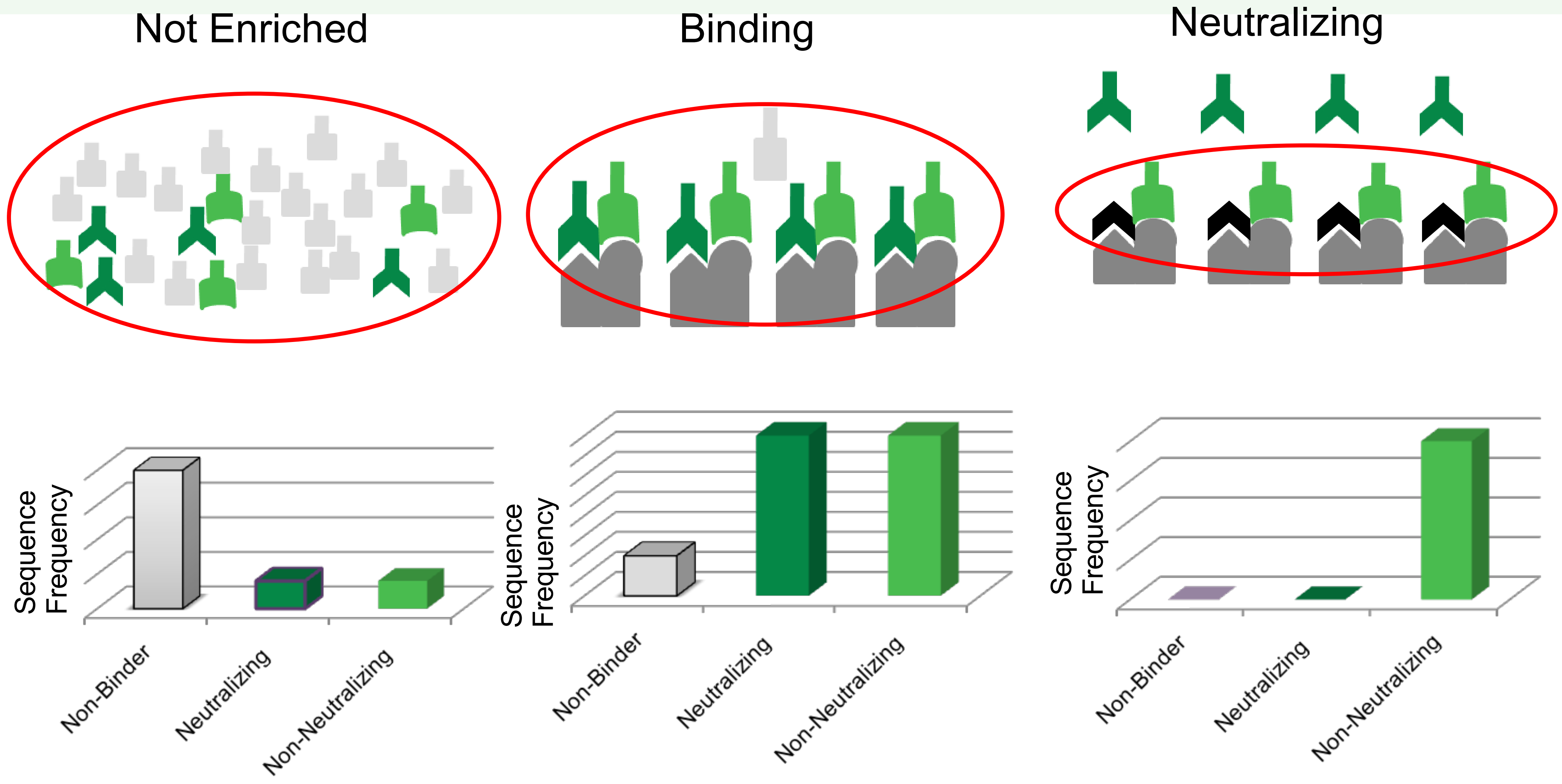
Rapid Versatile Affinity Selection Process



No cells, No Phage, No Vectors

X-BODY
BIOSCIENCES

Schematic of Sequence Frequency Predicting Binding Properties



Target protein

Active site ligand

Requirements for the Bioinformatics System

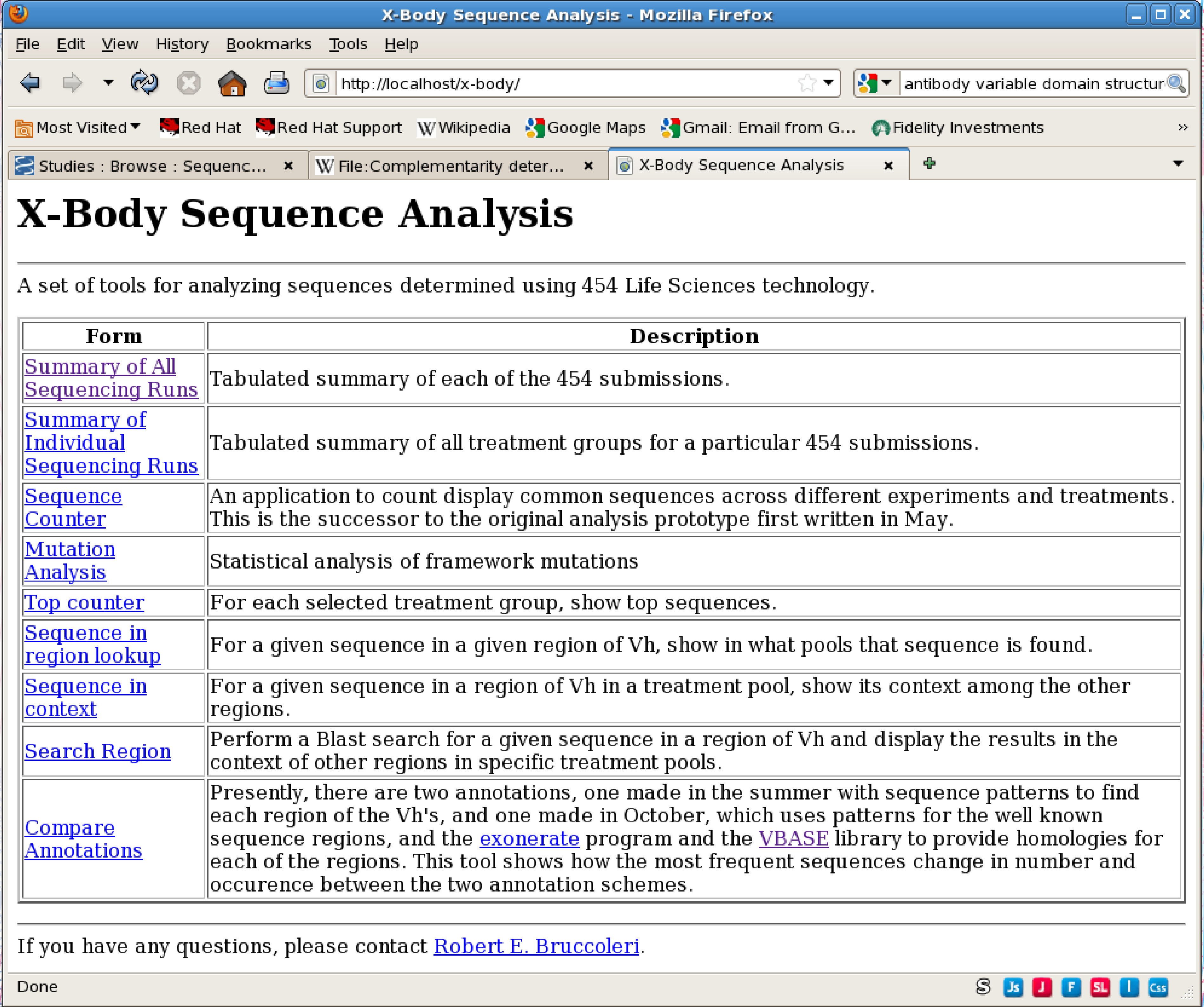
- Categorize sequences by experimental treatment
- Parse each sequence into Vh regions (CDR1, 2, and 3, FR 1, 2, 3 and 4)
- Provide analytic tools for X-Body scientists to identify sequence enrichment.
  - Implement tools quickly because new tools are driven by previous analyses.

Sequence Processing

- Each aliquot of the library is tagged with a unique barcode sequence.
- Reads start at the 3’ of the Vh gene and read in order: barcode, piece of C region, FR4, CDR3, FR3, CDR2, FR2, CDR1, FR1.
- Use Vbase definitions for frameworks and CDR1 and CDR2 to locate regions with perfect matches.
- Use position of FR3 and FR4 to locate CDR3.
- Use exonerate when perfect matches fail. Exonerate can handle frameshifts. Although CPU intensive, exonerate is used on a minority of reads because perfect matches work most of the time.
- Track and tabulate all results into a relational database.

Presentation of analysis

- Simple web applications using Perl CGI.
- Forms are driven by experiment design tables
    - Data selections based on targets or other experimental details.
- Displays are likewise organized by experimental design tables.
- Rich linking of data.
- Drill down to individual sequences



X-Body Sequence Analysis

A set of tools for analyzing sequences determined using 454 Life Sciences technology.

Form	Description
Summary of All Sequencing Runs	Tabulated summary of each of the 454 submissions.
Summary of Individual Sequencing Runs	Tabulated summary of all treatment groups for a particular 454 submissions.
Sequence Counter	An application to count display common sequences across different experiments and treatments. This is the successor to the original analysis prototype first written in May.
Mutation Analysis	Statistical analysis of framework mutations
Top counter	For each selected treatment group, show top sequences.
Sequence in region lookup	For a given sequence in a given region of Vh, show in what pools that sequence is found.
Sequence in context	For a given sequence in a region of Vh in a treatment pool, show its context among the other regions.
Search Region	Perform a Blast search for a given sequence in a region of Vh and display the results in the context of other regions in specific treatment pools.
Compare Annotations	Presently, there are two annotations, one made in the summer with sequence patterns to find each region of the Vh's, and one made in October, which uses patterns for the well known sequence regions, and the exonerate program and the VBASE library to provide homologies for each of the regions. This tool shows how the most frequent sequences change in number and occurence between the two annotation schemes.

If you have any questions, please contact Robert E. Bruccoleri.

Lessons Learned

- Sequencing must be consistent with experimental design. Here, 454 sequencing provides data on CDR3, most important region in the gene, and sequencing starts from 3’ end.
- Capture experimental design with the sequences.
- Use relational databases to keep track of all sequences. Greatly facilitates both processing (track success and failures) and analysis.
- Solve easy problems first and hard problems later.

Resources

- PostgreSQL (http://www.postgresql.org/)
- Exonerate (http://www.ebi.ac.uk/~guy/exonerate)
- Frescobi (watch http://search.cpan.org) - will be released when tidied up.

Acknowledgement

Eric Yang

Summary of Advice

- Close collaboration with the biologists is essential.
- Find and evaluate good tools.
- Use relational databases when appropriate.
- Capture experimental design in project database.
- Solve easy problems first and hard problems later.
- Be fearless – challenges of NGS can be vanquished!