A troubleshooting guide:

Sequencing experts give tips on how to assemble and align sequence data.

Sequence Assembly and Alignment

Genome Technology

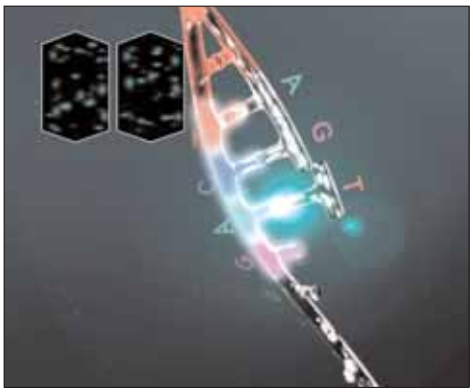
GenomeWeb

The GenomeWeb Intelligence Network

www.roche-applied-science.com

Genome Sequencer FLX System

Longer sequencing reads mean more applications.



Sequencing-by-Synthesis: Using an enzymatically coupled reaction, light is generated when individual nucleotides are incorporated. Hundreds of thousands of individual DNA fragments are sequenced in parallel.

In 2005, the Genome Sequencer 20 System was launched
- Read length: 100 bases
- 20 million bases in less than 5 hours

In 2007, the Genome Sequencer FLX System was launched
- Read length: 250 to 300 bases
- 100 million bases in less than 8 hours

Available in 2008, the Genome Sequencer FLX with improved chemistries
- Read length: >400 bases
- 1 billion bases in less than 24 hours

More applications lead to more publications

Proven performance with an expanding list of applications and more than 80 peer-reviewed publications.
Visit ***www.genome-sequencing.com*** to learn more.

454 LIFE SCIENCES

Roche Diagnostics
Roche Applied Science
Indianapolis, Indiana

Roche

For life science research only. Not for use in diagnostic procedures.

Table of contents

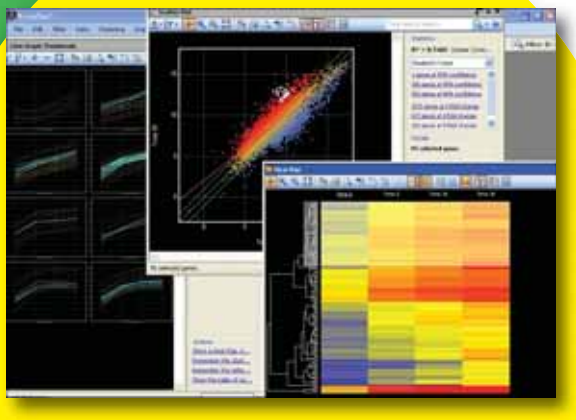
ArrayStar® shines at straightforward microarray analysis

If your studies involve gene expression, ArrayStar is the only microarray software that helps you do so much analysis so easily. Why shuffle between a confusing assortment of complex, expensive, hard-to-learn software when you can quickly and easily perform the analyses you want with a single user-friendly program?

And ArrayStar is as straightforward to own as it is to use. When you purchase ArrayStar, the perpetual license makes it yours forever.

Your research benefits from these features!

- Compatible with Affymetrix and NimbleGen normalized data files
- Import wizard supports .txt files
- Conducts a wide range of analyses and visualizations to evaluate expression patterns
- Identifies genes with similar expression patterns
- Clusters images to identify groups of co-regulated genes
- Plots gene expression levels to view relationships with Line Graphs and Thumbnails
- Displays expression levels of many genes across multiple experiments with a Heat Map



A selection of genes highlighted within simultaneous multiple views of ArrayStar

Affymetrix is a trademark of Affymetrix, Inc.
NimbleGen is a trademark of NimbleGen Systems, Inc.

To receive a 30-day fully-functional free trial version visit us at www.dnastar.com/arraystar

For more information, or to order ArrayStar today, please call toll-free in the **US, 1.866.511.5090** or in the **UK, 0.808.234.1643**

DNASTAR ArrayStar

www.dnastar.com/arraystar

DNASTAR, Inc • Madison, WI, USA • www.dnastar.com
Toll free: 866-511-5090 • 608-258-7420 • info@dnastar.com

Letter from the editor

At long last, you've got your sequence data. But now you have to figure out what it means. To make heads or tails of all that information, you might want to align and then assemble your genome of choice. Gone, though, are the days of manually post-processing the snippets of sequence. The flurry of activity as more and more genomes are sequenced (Dog! Honey bee! Shark!) has brought more advanced and accurate ways of aligning and assembling whatever your favorite genome might be. To help you slog through all the choices and decisions, *Genome Technology* has rounded up experts in the sequence alignment and assembly fields. These experienced veterans give their advice on how to approach sequence alignment, choose an assembly algorithm, and when to declare yourself "done!" with putting that genome's pieces together. Without further ado, but with much thanks to the following contributors, here are their thoughts on the questions facing researchers in sequence alignment and assembly.

— Ciara Curtin

Index of experts

Genome Technology would like to thank the following contributors for taking the time to respond to the questions in this tech guide.

Stephen Altschul
Senior Investigator
National Center for Biotechnology Information

Serafim Batzoglou
Assistant Professor
Stanford University

Gustavo Glusman
Senior Research Scientist
Systems Biology Institute

Xiaoqiu Huang
Associate Professor
Iowa State University

Sudhir Kumar
Professor
Arizona State University

Elliott Margulies
Investigator
National Human Genome Research Institute

Darren Platt
Informatics Department Head
DOE Joint Genome Institute

Mihai Pop
Assistant Professor
University of Maryland

How do you decide whether to take a global, local, or hybrid approach to sequence alignment?

If you know beforehand that two or more sequences are globally related, then it is almost always advantageous to use a global alignment algorithm to compare them. This will cut down on noise from random local similarities, and will force the alignment of related sequence regions that a local algorithm might leave unaligned. However, if it is not known whether the sequences being compared are globally related, or indeed whether they are related at all, then a local alignment algorithm is to be preferred. Because most database searches are exploratory in this sense, almost all popular sequence database search programs, such as FASTA and Blast, employ local alignment methods.

A few years ago, Dr. Yi-Kuo Yu described what he called a "hybrid" alignment method, which was a combination of the Hidden Markov Model approach to local sequence alignment (which considers all possible paths through a path graph) and the Smith-Waterman algorithm (which seeks only the optimal local alignment). This hybrid approach, which is a local alignment method, has significant advantages from a statistical perspective, and perhaps advantages from the perspective of sensitivity as well. However, there is so far no widely available software implementing this approach.

— Stephen Altschul

> "Almost all popular sequence database search programs, such as FASTA and Blast, employ local alignment methods."
>
> *— Stephen Altschul*

The global alignment model is appropriate in specific problems such as alignment of orthologous proteins or overlap of reads in DNA sequencing projects. Local alignment is best used for searching a database for a sequence of interest. Hybrid approaches are needed when aligning large genomic regions between different species.

— Serafim Batzoglou

Most sequence assembly tools start by identifying overlapping sequence reads and aligning them. For example, Phrap uses cross match as its alignment engine. Arachne implements its own fast heuristic to identify overlapping reads, by indexing subsequences (words); the optimal alignment is obtained using dynamic programming in a process conceptually related to the FASTA algorithm. In the context of sequence assembly, therefore, the choice of alignment algorithm is out of the user's hands, having already been addressed by the assembly tool developer.

— Gustavo Glusman

Start with a local alignment program named SIM. If the alignment covers most of the sequences, use a global alignment program named GAP. If the sequences contain similar regions separated by different regions, use a pair-wise alignment program named GAP3 or a multiple alignment program named MAP2.

— Xiaoqiu Huang

In global alignments, sequences are aligned beginning to end, with gaps inserted whenever needed. This procedure is appropriate for protein (amino) sequences of individual genes, ribosomal RNA sequences, and short stretches of the genomic sequence. For genomic DNA, it is always necessary to account for medium and large-scale rearrangements in addition to large sequence insertions and deletions, which necessitates the building of local alignments. For aligning DNA sequences of genome segments coding for proteins (protein-coding genes), one will often need to take a hybrid approach in which the first step is to align the translated protein sequences by taking a global approach, which is followed by the adjustment of exon DNA sequences to reflect the protein alignment, and then the use of local procedure for aligning homologous intron sequences.

— Sudhir Kumar

I typically align genomic sequences, so a hybrid approach works best for me. A global alignment alone doesn't work when the parts of the genome you're trying to align are in different segments and all shuffled up, so you need to first pre-process the data to figure out which sequences should be aligning in the first place.

— Elliott Margulies

"The overarching answer is usually: do the right thing."

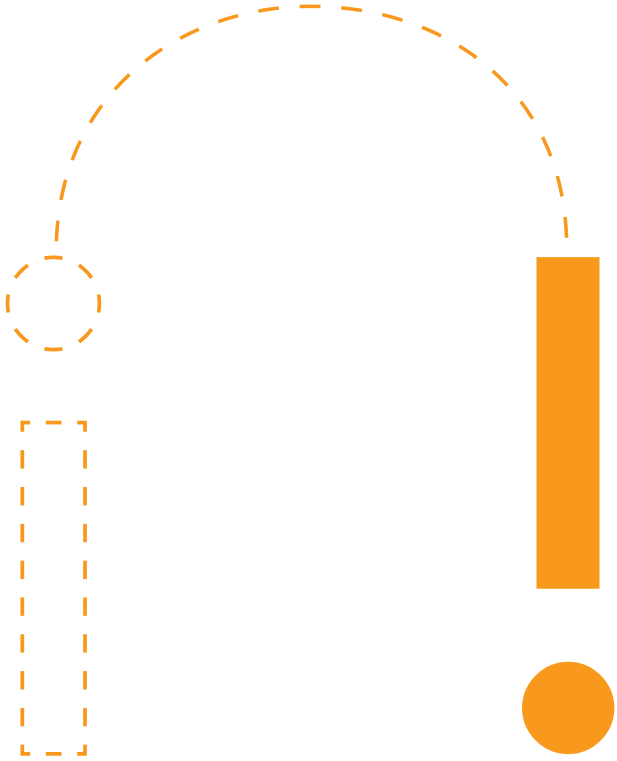
— Darren Platt

The overarching answer is usually: do the right thing. If you expect the sequence comes from a gene, then it should align in full over the coding area that you want put it in. So that typically means global alignment. In assembly, that's what you want. You want every read to align along the genome. In practice, you might put low-quality sequence off the ends and make sure it actually aligns properly.

— Darren Platt

In general I prefer to use a local alignment procedure, followed by post-processing in case I need to place additional constraints on the alignment. For example, when mapping a sequencing read to a reference genome, I try to ensure the alignment spans the entire read (i.e. a global alignment from the point of view of the read). Requiring a global alignment for the reads, however, would miss polymorphisms between the DNA being sequenced and the available reference.

— Mihai Pop



Turn sequencing on its head.

It's here. Now you can perform a range of genomic experiments that you only dreamed of before. The Illumina Genome Analyzer changes the way you approach every kind of genomic experiment. The discoveries are happening, the wait is over, and our customers are publishing their results.

Change the way you think about sequencing.
Join the growing Illumina Community.

Do more with sequencing today:

www.morethansequencing.com

illumina®

What scoring function do you use? Do you allow gaps?

Most alignment methods profit by allowing gaps. One important exception, perhaps, are multiple local alignment algorithms which seek, many copies of short sequence motifs in DNA or protein sequences. These algorithms may eschew gaps for algorithmic efficiency, to reduce random noise, and/or because most of the motifs sought do not contain gaps. However, standard pair-wise alignment algorithms essentially all allow gaps.

In the context of pair-wise local or global sequence comparison, there is no single "optimal" scoring system. It has long been recognized that different amino acid and DNA substitution matrices are tailored to different evolutionary distances so that, for example, there is not a single PAM or a single BLOSUM matrix, but rather a series of them. Again, the more one knows a *priori*, the better one can do. For example, if the two protein sequences being compared are very closely related, then a matrix with high "relative entropy" (such as a low-numbered PAM or a high-numbered BLOSUM matrix) should be used. The default matrices employed by FASTA and Blast have been chosen to do well at detecting sequence similarities at the borderline of statistical significance — the so-called "twilight zone." However, very short queries require more information per alignment position to rise above background noise, and thus call for higher relative-entropy matrices.

> "Alignment, by definition, includes gaps. Every alignment problem has its own scoring function or set of scoring functions, and defining those is still an open problem."
>
> *— Serafim Batzoglou*

If one has a multiple alignment and wishes to search for additional related sequences, then scoring systems can be tailored to be position-specific, an approach taken by PSI-Blast and many Hidden Markov Model methods. Also, recent work has suggested that it is often profitable to modify standard substitution matrices when comparing sequences with biased amino acid or nucleotide compositions.

It has been recognized for many years that "affine" gap costs, which penalize for the existence of a gap as well as incrementally for its length, are generally superior to "linear" gap costs, which charge only for each residue inserted or deleted. Gap costs need to be chosen in tandem with a specific substitution matrix, and default gap costs for database search programs have usually been optimized by trial and error. HMM approaches can use position-specific gap costs as well as position-specific substitution costs. Also, it is possible to define pair-wise alignment gap costs that allow an algorithm to skip, or decline to align, regions in both sequences simultaneously.

— Stephen Altschul

Alignment, by definition, includes gaps. Every alignment problem has its own scoring function or set of scoring functions, and defining those is still an open problem. Lately, machine learning techniques

have been introduced to the problem of scoring protein alignments.

— Serafim Batzoglou

When comparing divergent sequences, the choice of parameters can significantly affect the results. Relevant parameters include the scoring matrix, the gap opening, and extension penalties, etc. In the context of sequence assembly, the sequences being compared are usually almost identical, differing only by sequencing errors and/or by polymorphisms. The specific parameters for sequence alignment are typically optimized by the tool developer. Typically, there is no need to modify such parameters, particularly for large, high-throughput projects. Some packages like Sequencher allow for more fiddling with assembly parameters, which can be important for assembling sequences with high polymorphism rates.

— Gustavo Glusman

Gaps are allowed.

For protein sequences, select a substitution matrix based on the alignment percent identity: Use BLOSUM50 if percent identity < 30%, BLOSUM62 if 30% <= percent identity <= 45%, and BLOSUM100 if percent identity > 45%.

Use a gap open penalty of 10 with BLOSUM50, 14 with BLOSUM62, 18 with BLOSUM100. Use a gap extension penalty of 2 in each case.

For DNA sequences, use a match score of 10, a mismatch score of -18, a gap open penalty of 60, and a gap extension penalty of 2.

— Xiaoqiu Huang

All alignments need to allow for gaps, because the alignment is the process of inserting gaps in sequences in order to estimate the base homology muddled by the insertion-deletion and substitution processes. The difference is their contribution to the alignment score creates differences between local and global alignments. In the local alignments, focus is on shared regions of high similarity and regions that do not show high sequence homology between sequences are simply ignored. In the global alignments, there are penalties for inserting a gap and for extending it to span multiple bases. As for the alignment parameters used for scoring the cost of alignments, the relationship between the alignment accuracy and the gap and substitution penalties is not straightforward. For example, the use of commonly employed alignment parameters leads to only slightly worse alignments than those obtained where we to know the true penalties for inserting gaps and allowing base substitutions.

— Sudhir Kumar

The scoring function depends on how diverged the genomic sequences are that I'm trying to align. I'll use a less stringent matrix for more diverged pairs of species and a more stringent matrix for more closely related species. I definitely allow for gaps. Genomic sequence alignments would be virtually impossible with allowing for gaps.

— Elliott Margulies

The only reason you should get a mismatch or a gap would be if there is an actual error in the sequence read, otherwise it should match the genome or if there's a SNP in your sequence *(continued on p.19)*

Biocomputing systems that analyze the environment without warming it.

High-Performance Biocomputing Solutions

The CodeQuest™ and DeCypher® systems from TimeLogic® handle huge genomic analysis projects faster than most clusters, yet use less electricity than a standard light bulb.

Both solutions employ the DeCypher Engine™ accelerator cards to deliver the performance of hundreds of CPUs, so you can:

- Map genomic signatures, SNPs and microarray probes with high specificity using Tera-Probe™
- Match millions of nucleic sequences to nt with Tera-BLAST™
- Compare proteins to UniProt using Smith-Waterman
- Identify protein families using hmmpfam searches

Keep your genomics projects on track and your computing budgets above water. Contact TimeLogic for a free performance metric today!.

CodeQuest

Accelerated workstation for your lab.

DeCypher

Scalable Solutions for the Enterprise.

North America
Active Motif, Inc.
Toll Free 877-222-9543 x 4
tlsales@activemotif.com

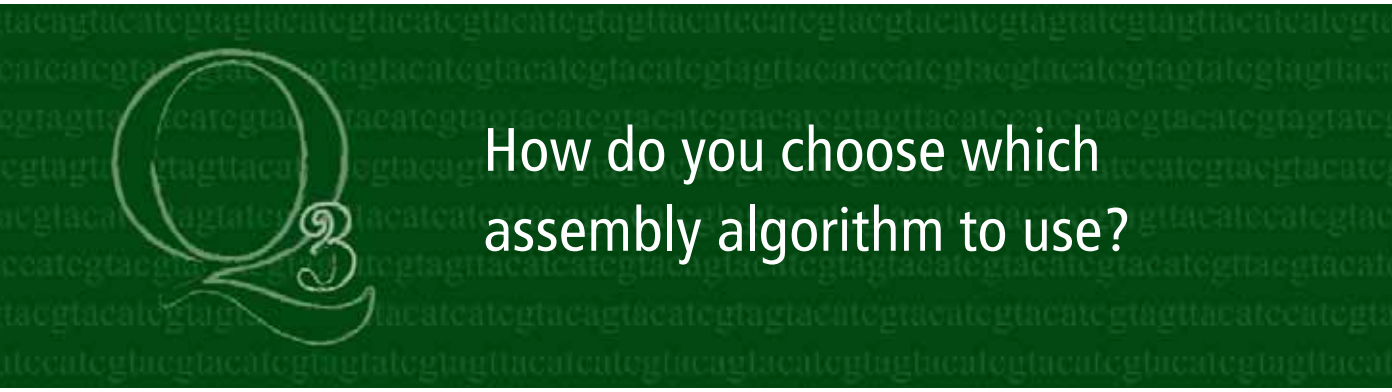
Europe
Active Motif Europe
Direct +32 (0)2 653 0001
tlsales@activemotif.com

Asia
Active Motif Japan
Direct +81 (0)3 5225 3638
japantech@activemotif.com

Additional international distributors: www.timelogic.com/distributors

TimeLogic®
biocomputing solutions

A brand of Active Motif,® Inc.



How do you choose which assembly algorithm to use?

The cutoff threshold for calling two reads "overlapping" depends on the read error rate, which in turns depends on (1) the sequencing technology and (2) the amount of successful error-correction done computationally by aligning the reads to each other. The cutoff threshold is chosen heuristically so that the majority of true read overlaps are accepted. Regarding the choice of assembly algorithm: for complicated sequencing projects, applying the existing assembly systems is a significant project, and therefore often the assembly algorithm development team is the one who applies their algorithm to a newly sequenced genome.

— Serafim Batzoglou

Be pragmatic: use the assembly tool that best fits your workflow and working environment, the tool that peers in your institution use, and for which you'll have support. If, for example, your institution has an established data pipeline using Phrap, you're better off just using it.

— Gustavo Glusman

"Be pragmatic: use the assembly tool that best fits your workflow and working environment."

— Gustavo Glusman

- Use Phrap for assembly of BAC clones.
- Use CAP3 for assembly of EST sequences.
- Use the parameter cutoffs -p 90 and -d 100 with CAP3 for conservative assembly.
- For whole-genome assembly, use PCAP with the default parameter cutoffs.

— Xiaoqiu Huang

Most of the data I analyze is already assembled. I leave it to the assembly pros to make that work well. On occasion, when I've had to assemble genomes from low-redundancy sequencing efforts, I've turned to Phusion from Jim Mullikin.

— Elliott Margulies

With assembly algorithms I'm biased because I developed one, so I tend to use Forge which is the one I've been working on myself for about seven years. I've seen comparisons of the algorithms over the years. I think, generally, Arachne comes out on top in those competitions. Forge wasn't entered in that particular competition. What I've noticed is in practice people use what's actually most useable — if it's what you can get to run on your machine. People still use Phrap, even though it was first published in 1994 or so, because it is really useable. I think people have to weigh the ease of use against the theoretical accuracy of the algorithm and the initial comparison. There isn't a lot of comparison because it is very hard to do. In terms of the alignment algorithms, I think people are probably looking at trade-off between speed and sensitivity there. I don't think they're mature enough for people to know which is going to be best to align a bunch of reads. I think algorithms that can take into account gene array data is central.

— Darren Platt

(continued on p.19)

Even in reverse the advantage is clear.

Easy-to-see and easy-to-use all in one box. We've developed the ultimate in QRT-PCR convenience and performance by combining the new Thermo Scientific Verso™ RT enzyme with Thermo Scientific ABsolute™ Blue QPCR Master Mixes.

- Reproducible - Easy-to-see inert blue dye minimizes operator pipetting errors
- Fast - RT Enhancer eliminates the need for additional DNase I treatment step
- Flexible - Alternative RNA template priming options to maximize sensitivity
- High Performance - Based on the proven range of ABsolute™ QPCR Master Mixes

For the complete offering of ABsolute™ QPCR & QRT-PCR master mixes visit: www.thermo.com/abgene

Thermo Scientific Verso™ QRT-PCR Kits
Optimized mixes are available in both 1-Step and 2-Step kits for all QPCR machines for any application.

Part of Thermo Fisher Scientific

Thermo SCIENTIFIC

What approach do you use to deal with very short reads?

New algorithms are being developed for assembly with short reads. Solexa has an algorithm that accompanies their system for mapping short reads of a resequencing project onto the target genome. We have a recent paper in PloS One (Sundquist *et al*.) proposing a whole-genome *de novo* sequencing and assembly protocol with short unpaired reads.

— Serafim Batzoglou

Genomic sequencing using short reads is an area under very active development and in which protocols have not yet been firmly established. An article of particular interest was recently published by Sundquist *et al*. describing a hierarchical protocol for whole-genome sequencing based on short reads.

— Gustavo Glusman

Long reads of low coverage are used to help place short reads of high coverage.

— Xiaoqiu Huang

I've just started to look at short-read data and find that a new program from Ewan Birney's group called Velvet does quite well.

— Elliott Margulies

That's probably the toughest area right now that everybody's so excited about. The major, major issue is most algorithms rely on CANA or Merle some contiguous identical sequence to start the search process, i.e. 14 or 17 bases are identical to a reference genome. A single error inside a 25-base pair read won't be alignable. You have to have hash-based algorithms you can use and allow for gaps and mismatches in the keys. Basically, split-key search methods are really essential if you are going to align a read of 25 bases or less if there's an error in there potentially.

— Darren Platt

For aligning short reads to a reference genome I have had pretty good success using both MUMmer and Vmatch. The alignment parameters must strike a balance between sensitivity and specificity: requiring high-fidelity alignments dramatically reduces the time needed to perform the alignments, yet might miss regions of polymorphism.

— Mihai Pop

How do you ensure high-quality assemblies? What's your process for detecting errors?

There are many heuristic methods for detecting assembly errors. Two general methods of evaluation are simulation, and comparison of portion of an assembled genome with an independently derived sequence that we believe was assembled correctly. On top of that, there are a great variety of sanity checks that can be done on an assembly, such as checking for consistency with independently derived mate pairs and other mapping data.

— Serafim Batzoglou

Visualizing the results of the assembly is crucial for quality control. This can be done using software like Consed, Sequencher, etc. Looking at the assembly overview, are there regions with a surprisingly high coverage? This may imply an over-collapsed repeat in the sequence. Zooming into the contigs, are there clusters of ambiguities in regions with sufficient coverage? This may indicate a misassembly.

— Gustavo Glusman

Examine assembly regions with a sufficient number of unsatisfied read pairs. Check assembly results against physical maps.

— Xiaoqiu Huang

A large portion of sequence data I analyze comes from the NISC Comparative Sequencing Program, where they sequence and assemble BAC-based sequences. While some automated tools exist to put these sequence data correct most of the time, we go through a manual curation process. This is important since these data are used as a "gold standard" against which other genome assemblies are compared.

— Elliott Margulies

To some extent there is no correct answer. The best proxies for quality seem to be clone coverage and read coverage which — if you've got a consistent clone coverage across the region, all the clones, the pairs are not stretched or compressed and there's a reasonable depth — then you've probably got a reasonably good assembly, structurally. If you look for areas with only a single read covering a particular region of assembly, that's almost always a mistake. So that's globally looking at assembly and making sure everything's in the right place. The next thing is at the base-pair level for the consensus. With these new technologies, you have so much depth of coverage so that every area should be covered with 10 or 20 reads. Most interesting bases to inspect are the ones where there is controversy. So if you don't get 90 percent of the reads giving you one answer or 50-50, if it is a diploid organism, you've got reason to go and take a look at it. Unfortunately, these technologies do make systematic errors. You can get quite controversial bases even if you've got fifty-fold coverage — which is something we're are all going to have to deal with eventually.

— Darren Platt

For Sanger data (or any other data-set providing mate-pair information), we use a combination of automated diagnostic tools (the amosvalidate tool from the AMOS package) and manual inspection of the assembly in Hawkeye. The automated pipeline examines the assembly to detect violations of the mate-pair constraints (mis-oriented, stretched, or compressed mate-pairs), high-quality discrepancies between co-assembled reads, as well as breakpoints in the

(continued on p.19)

COGENICS™
A DIVISION OF CLINICAL DATA

Comprehensive Pharmacogenomics and Molecular Biology Services

Why spend valuable research resources recreating or maintaining systems in-house? Cogenics genomics services offer cost-effective and precise alternatives for fast turn-around solutions that respect your valuable discovery resources.

Whether your projects are large or small, basic research-oriented or require FDA submission, Cogenics has service solutions for you.

http://www.cogenics.com/gogreen

USA: 1-877-226-4364
UK: +44 (0)1279-873837
France: +33 (0)456-381102

E-mail: sales@cogenics.com

How do you close the gaps? At what point is a genome considered "finished"?

The method of choice for closing gaps strongly depends on the resources available to the scientist. If one has easy access to a sequencing facility, adding more shotgun reads may be the most cost-effective method for overall quality improvement, and it will eventually close most gaps. Some regions are hard to sequence (e.g. regions with homopolymers) and if changing the sequencing chemistry is an option, this may be the solution. Finally, there's the directed method of designing PCR primers, amplifying and sequencing the products in the hope these will bridge the gaps (or to improve regions of low quality).

As for when to consider the sequence "finished," again, some very pragmatic considerations: First of all, keep in mind that there is polymorphism in the population; it doesn't make a lot of sense to require a lower error rate in the sequence than the polymorphism rate. Second, does the resulting sequence cover all "interesting" areas with high enough quality? If new areas of interest are identified, those regions can be resequenced later to achieve higher quality, as needed. Finally, can one locate most known mRNAs and ESTs in the resulting genomic sequence? Can it account for all previously known sequence information about the genome (or locus, if sequencing a specific region)?

— Gustavo Glusman

> "The method of choice for closing gaps strongly depends on the resources available to the scientist."
>
> *— Gustavo Glusman*

Use Consed/AutoFinisher to aid in gap closures. The genome is considered finished when biological questions can be addressed by using the genome.

— Xiaoqiu Huang

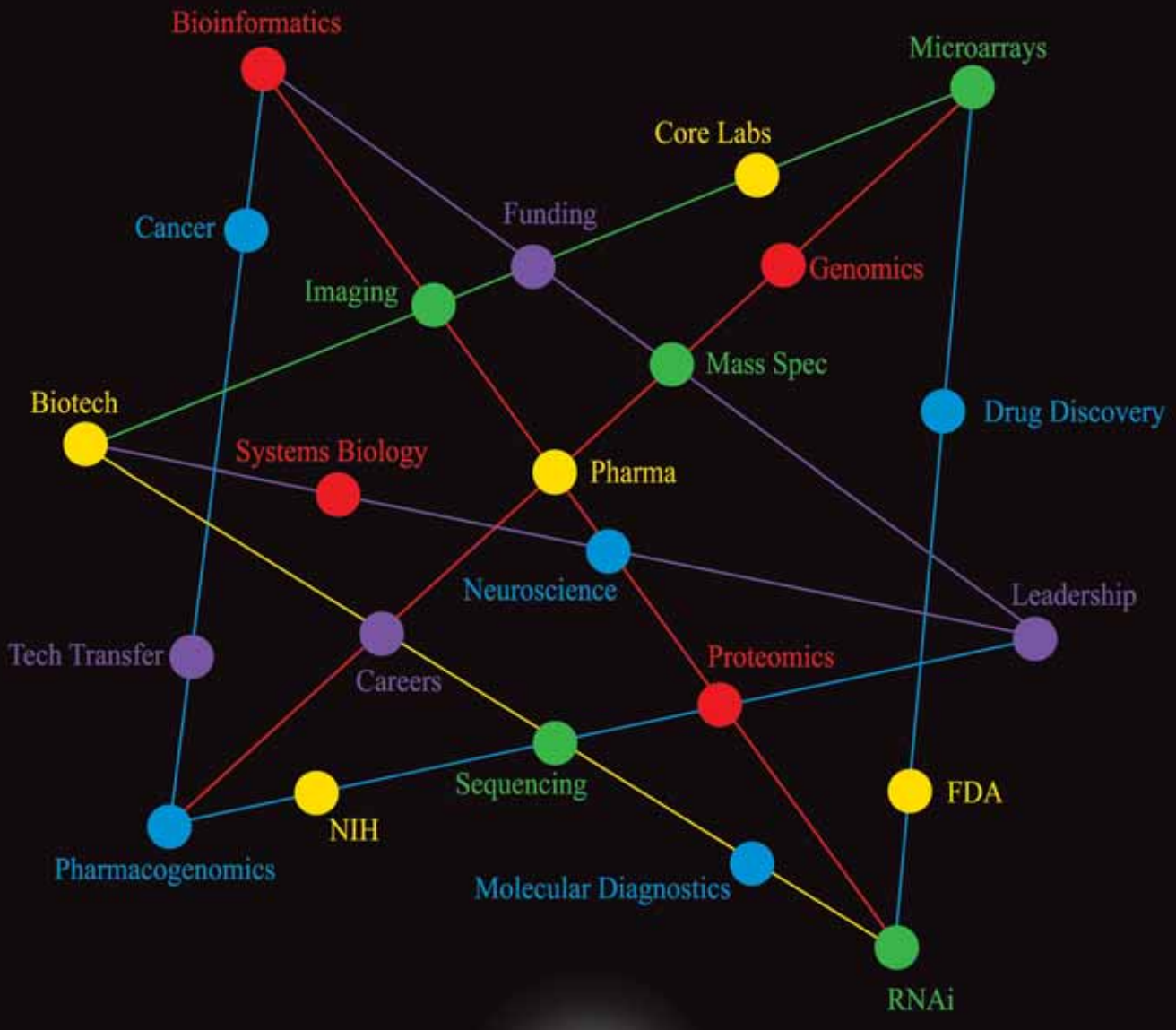
When dealing with BAC-based sequence assemblies, we typically screen BAC libraries three or more times before giving up. We are beginning to find that many of these gaps are due to low representation in the BAC library. However, sometimes biology is to blame — this can be determined from aligning the sequences to other species in order to get a better idea of what is happening throughout evolution.

— Elliott Margulies

At JGI, the process is quite extensive. They have various technologies at their disposal: clone walking and primer walking into a gap, building a shadow library of a clone extends into a gap, multiplex PCR if you don't know the orientation of the contigs. We are using 454 to fix gaps and assemblies and now we are using Solexa to fix gaps in our 454 assemblies in controversial areas. In terms of knowing when to finish, it's a pragmatic decision still. If it's a bacterium, it should go into a single piece and it's not until we've exhausted every known chemistry at our disposal that we'll officially give up on closing a gap — and that happens still. If it's a fungus, then it gets a little bit harder to contemplate closing all of the gaps. It's usually possible *(continued on p.19)*

The GenomeWeb Intelligence Network.

Connecting the dots for researchers worldwide.



GenomeWeb

Q2: What scoring function do you use? Do you allow gaps? (continued from p.10)

relative to the reference genome. The main algorithms I've worked on can actually use the quality of each base to determine the likelihood of there being a gap. You actually vary the penalty on a position by position basis. The other further complication is you need to take into account the technology that was used to do the sequencing. If it was a 454, it is very likely to create an insertion or a deletion and very unlikely to have a substitution; if it's a Sanger read you have equal probability of a base being an addition, substitution, or deletion.

— Darren Platt

I run most alignments using the MUMmer package (specifically nucmer). This program implements a gapped alignment algorithm with affine gaps very similar to the default alignment used in Blast and FASTA.

— Mihai Pop

Q3: How do you choose which assembly algorithm to use? (continued from p.12)

For typical assembly tasks (e.g. assembly of a bacterial genome with Sanger data) we generally use Celera Assembler, though any of the large-scale assemblers (e.g. Arachne, PCAP) should also be suitable. My primary reason for using Celera Assembler is familiarity with the code, which makes it easier to assess the quality of the assembly and tweak its parameters. For small molecules (genes, viruses, etc.) we generally use Minimus (developed in our group), while for 454 data we use the newbler assembler.

For small assemblies (bacterial genomes or smaller), we often run multiple assemblies with different parameters (e.g. different similarity cutoffs) and choose the assembly that produced best results. For large eukaryotes, we generally use parameters that have worked well in the past.

— Mihai Pop

Q5: How do you ensure high-quality assemblies? What's your process for detecting errors? (continued from p.15)

alignment of singleton/shrapnel reads to the assembly. Such features often correlate with mis-assemblies.

— Mihai Pop

Q6: How do you close the gaps? At what point is a genome considered "finished?" (continued from p.17)

if the organism is well-behaved. Once it gets above about 15 megabases, we euphemistically go to "genome improvement" to order primers to fill all the potential gaps, but if for whatever reason we don't succeed, we don't necessarily go in there and try to plug every single one of them.

— Darren Platt

We find that many gaps can be automatically closed by relaxing the trimming constraints on the reads at the ends of contigs, as the trimming of poor-quality data from the reads is (by necessity) overly conservative.

In our research, we focus primarily on in silico finishing. In this context, "finished" implies a state of the genome assembly that is most consistent with the sequence and mate-pair data. As much as possible, the contigs should be ordered and oriented in a scaffold and any mis-assemblies corrected. This definition of "finished" provides most information for downstream comparative analyses and represents a good substrate for the design of targeted experiments aimed at completely closing all the gaps in the assembly.

— Mihai Pop

Plant & Animal Genome XVI

The International Conference on the Status of Plant & Animal Genome Research

January 12-16, 2008

Town & Country Hotel, San Diego, California

Speakers

Jerry Caulder, Finistere Partners, LLC, USA
David Baulcombe, John Innes Institute, UK
Gilbert Omenn, University of Michigan, USA
Susan McCouch, Cornell University, USA
Michael Ashburner, EMBL, European Bioinformatics Institute, UK
Eddy Rubin, DOE, Joint Genome Institute, USA
Steve Jacobson, UCLA, USA
Steve Horvath, UCLA, USA

Workshops

- Abiotic Stress
- Allele Mining
- Apomixis Aquaculture
- Banana (Musa) Genomics
- Barley
- Bioinformatics
- Brassicas
- Brachypodium Distachyon
- Cattle/Sheep
- Challenge Program: Unlocking Crop Genetic Diversity for the Poor
- Citrus Compositae
- Computer Demonstrations
- Cotton
- Compositae
- Connectrons Cool Season Legumes
- Cururbit
- Equine
- Forage & Turf Plants
- Forest Trees Fruit and Nut Crops
- Functional Genomics
- Host Pathogen Interactions
- Insect Genetics
- ICSB
- ICGI
- Int'l Grape Genome Project
- Int'l Lolium Genome Initiative
- IGGI
- ITMI
- Large-Insert DNA Libraries and Their Applications
- Legumes
- Maize
- Microarray Analysis
- Molecular Markers for Plant Breeders
- Mutation Screening
- NRSP-8
- NC 1010 Development and Implementation of Ontologies in the Database
- Organellar Genetics
- Plant Cytogenetics
- Plant Development and Signal Networks
- Interagency Working Group on Plant Genomics
- Plant Interactions with Pests and Pathogens
- Plant Transgene Genetics
- Plant Reproductive Genomics
- Polyploidy
- Poultry
- Proteomics
- QTL Cloning
- Reduced-representation
- Sequencing Methods and Applications
- Rice
- Rice Blast
- Root Genomics
- Solanaceae
- Sorghum and Millets
- Soybean Genomics
- Statistical Genomics
- Sugar Beet
- Swine
- Swine Genome Sequencing
- TAIR
- Weedy and Invasive Plant Genomics

Organizing Committee

CHAIRMAN:
Stephen R. Heller, NIST, USA

PLANTS:
Michael Gale, John Innes Center, UK
Ed Kaleikau, USDA/CSREES, USA
Dave Matthews, USDA, ARS Cornell University, USA
Graham Moore, John Innes Center, UK
Jerome P. Miksche, Emeritus Director, USDA Plant Genome Program, USA
Rod Wing, University of Arizona, USA

ANIMALS:
Mary Delany, UC-Davis, USA
Michel Georges, University of Liege, Belgium
Joan Lunney, USDA/ARS USA
Jim Reecy, Iowa State University, USA

ABSTRACT COORDINATORS:
Victoria Carollo, USDA, ARS, WRRC, USA
Gerard Lazo, USDA/ARS/WRRC, Albany, CA, USA
David Grant, USDA/ARS & Iowa State University, USA

Sponsors

USDA, Agricultural Research Service
USDA, National Agricultural Library
USDA, NRI Competitive Grants Office
USDA,Cooperative State Research, Education, and Extension Service (CSREES)
John Innes Centre
NCGR, National Center for Genome Resources

Presented by

Scherago International
525 Washington Blvd., Ste. 3310
Jersey City, NJ 07310
201-653-4777 x20
fax: 201-653-5705
E-mail: pag@scherago.com

For complete details, including on-line registration, visit our website at www.intl-pag.org

List of resources

When your sequence alignment and assembly problems go beyond what our experts have discussed here, the following publications and websites might come in handy. After all, that's where our experts turn when they have questions.

Websites

Arachne
http://www.broad.mit.edu/wga/

Blast
http://www.ncbi.nlm.nih.gov/BLAST/

CAP3
http://pbil.univ-lyon1.fr/cap3.php

Ensembl Genome Browser
http://www.ensembl.org/index.html

FASTA
http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi

Hawkeye
http://amos.sourceforge.net/hawkeye/

Minimus
http://amos.sourceforge.net/docs/pipeline/minimus.html

MUMmer
http://mummer.sourceforge.net/

Phred, Phrap, and Consed
http://www.phrap.org/phredphrapconsed.html

Sequencher
http://www.genecodes.com/

SIM
http://www.expasy.ch/tools/sim-prot.html

Vmatch
http://www.vmatch.de/

Publications

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou A. (2003). **Glocal alignment: Finding rearrangements during alignment.** *Bioinformatics*. 19(S1):i54-i62.

Carraro DM, Camargo AA, Salim AC, Grivet M, Vasconcelos AT, Simpson AJ. (2003). **PCR-assisted contig extension: stepwise strategy for bacterial genome.** *Biotechniques*. 34:626-8, 630-2.

Eddy, SR. (2004). **Where did the BLOSUM62 alignment score matrix come from?** *Nature Biotechnology*. 22:1035-1036.

Edgar RC, Batzoglou S. (2006). **Multiple sequence alignment.** *Current Opinion in Structural Biology.* 16:368-373.

Kumar S, Filipski A. (2007). **Multiple sequence alignment: In pursuit of homologous DNA positions.** *Genome Research.* 17:127-135.

Mullikin JC, Ning, Zemin. (2003). **The Phusion assembler.** *Genome Research.* 13(1): 81-90.

Sommer DD, Delcher AL, Salzberg SL, Pop M. **Minimus: a fast, lightweight genome assembler.** *BMC Bioinformatics.* 8:64.

Sundquist A, Ronaghi M, Tang HX, Pevzner P, Batzoglou S. (2007). **Whole-genome sequencing and assembly with high-throughput, short-read technologies.** *PLoS ONE* 2(5): e484.

Huang X. (1994) **On Global Sequence Alignment.** *Computer Applications in the Biosciences* 10, 227-235.

Huang X, Madan A. (1999). **CAP3: A DNA Sequence Assembly Program.** *Genome Research*. 9: 868-877.

Huang X, Wang J, Aluru S, Yang SP, Hillier L. (2003). **PCAP: A Whole-Genome Assembly Program.** *Genome Research*. 13: 2164-2170.

Mouse Genome Sequencing Consortium. (2002).

Initial sequencing and comparative analysis of the mouse genome. *Nature*. 420: 520-562.

Mullikin JC, Zemin N. (2003). **The Phusion Assembler**. *Genome Research*. 13: 81-90.

Schatz MC, Phillippy AM, Shneiderman B, Salzberg SL. (2007). **Hawkeye: an interactive visual analytics tool for genome assemblies.** *Genome Biology.* 8: R34.

Books

Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology
By David Gusfield
(May 1997) Cambridge Press; ISBN: 0521585198

Bioinformatics: Sequence and Genome Analysis
By David W. Mount
(July 2004) Cold Spring Harbor Laboratory Press; ISBN: 0879696877

Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acid
By Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison
(May 1998) Cambridge University Press; ISBN 0521620414

Computational Methods in Molecular Biology
Ed. by Steven Salzberg, David Searls, and Simon Kasif
(June 1998) Elsevier Science; ISBN: 0444828753

An Introduction to Bioinformatics Algorithms
By Neil C. Jones and Pavel A. Pevzner
(August 2004) MIT Press; ISBN: 0262101068

Introduction to Computational Molecular Biology
By Carlos Setubal and Joao Meidanis
(January 1997) PWS Publishing Company; ISBN: 0534952623

Upcoming Conferences

Exploring Next Generation Sequencing: Applications and Case Studies
October 17-18, 2007; Providence, RI

ESF-EMBO Symposium: Comparative Genomics of Eukaryotic Microorganisms: Eukaryotic Genome Evolution
October 20-25, 2007; Sant Feliu de Guixols, Spain

7th Cold Spring Harbor Laboratory/Wellcome Trust Conference on Genome Informatics
November 1-5, 2007; Cold Spring Harbor, NY

6th Georgia Tech-Oak Ridge National Labs International Conference on Bioinformatics, in silico Biology: Gene Discovery and Systems Genomics
November 15-17, 2007; Atlanta, GA

RECOMB Satellite Conference on Systems Biology
November 30- December 1, 2007; San Diego, CA

RECOMB 2008: 12th Annual International Conference on Computational Molecular Biology
March 30- April 2, 2008; Singapore

GME2007

Genomes, Medicine, and the Environment Conference

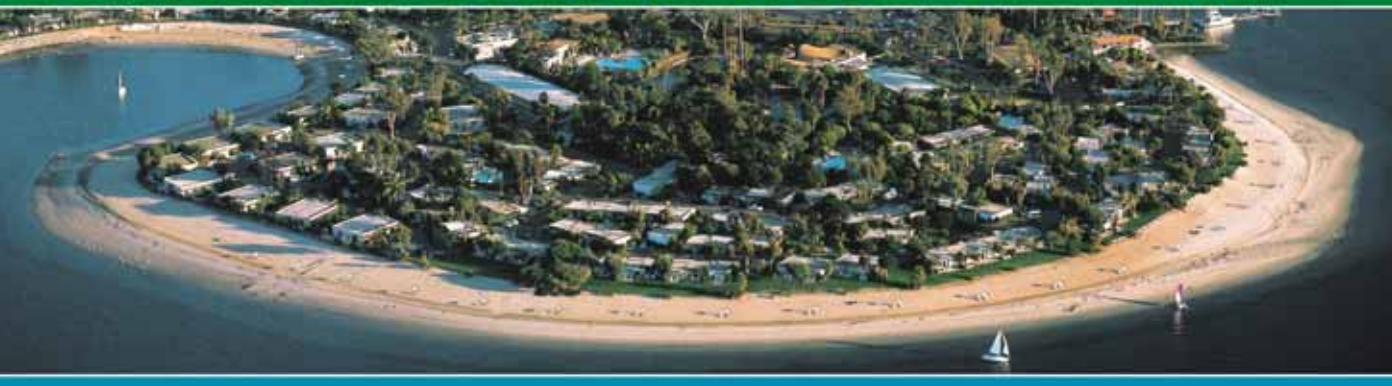
Paradise Point Resort & Spa, San Diego, CA • October 8 – 10, 2007

Please Join Us

› The J. Craig Venter Institute invites you to the Genomes, Medicine, and the Environment 2007 Conference. This three-day intensive meeting features presentations by world renowned genomics researchers. Meet experts in a variety of genomics disciplines, participate in interactive, in-depth discussions on future innovations in the field, and network with colleagues and life sciences suppliers.

Topics

- Synthetic Biology
- Biological Energy
- Environmental Genomics
- MetaGenomics
- Marine Microbiology
- Human Metagenomics
- Infectious Diseases
- Personalized Genomics
- Human Genomics
- Emerging Technologies



Paradise Point Resort & Spa, San Diego, CA

Call for Abstracts

› Poster abstracts are now being accepted! Do not miss the opportunity to present your research and discoveries. One abstract will be selected for oral presentation at the conference. Visit **www.jcvi.org/gme** for more information.

Register Now!

› Avoid on-site fees. Register now to save. Visit **www.jcvi.org/gme** for more information.

SPONSORS:

Roche illumina Applied Biosystems Genome Research Bentham Science Current Genomics Hamilton

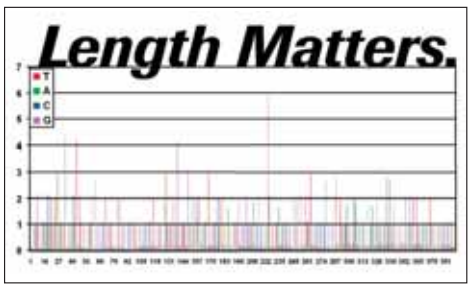
HOSTED BY:

J. Craig Venter Institute

www.roche-applied-science.com

Genome Sequencer FLX System

Longer sequencing reads mean more applications.



Flowgram showing a single read of 256 bases.
Each bar represents a discrete base (A, T, C, or G), and the height of a bar correlates to the number of bases in a specific position.

In a single instrument run, the **Genome Sequencer FLX System** generates over 400,000 reads of 200 to 300 bases with 99.5% accuracy per read.

- Perform *de novo* sequencing of whole genomes.
- Analyze full-length cDNA, including splice variants.
- Discover viral subtypes (*e.g.*, HIV).
- Uncover the diversity in metagenomic samples.

More Flexibility, More Applications, More Publications

Visit ***www.genome-sequencing.com*** to learn about the expanding number of peer-reviewed publications appearing weekly.

454 LIFE SCIENCES

Roche Diagnostics
Roche Applied Science
Indianapolis, Indiana

Roche

For life science research only. Not for use in diagnostic procedures.