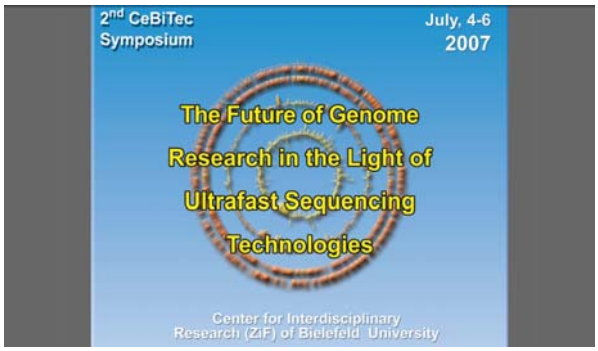


Overview

- Two Base Encoding in SOLiD
  - Michael Rhodes, Applied Biosystems
- Assembly of short read data
  - D.Evers, D.Platt (Forge)
  - Heinz Himmelbauer, MPI MG, (SHARCGS)
- Metagenomics / Microbes
  - Huson, Tübingen (MEGAN)
  - Stoye, Bielefeld
  - Richardson, JGI
- Exploring Transcriptomes
  - Weber, Düsseldorf

Two Base Encoding in the SOLiD Sequencing System

- Supported Oligo Ligation Detection
- *Massively parallel* sequencing by stepwise ligation with dye-labeled oligonucleotides
- up to 3 GB per run
- mate-pair analysis (25 bp)
- Complex genomes (mammalian)
- SNPs, CNVs, inversions, insertions, deletions
- 2-base encoding: calls each base twice *(without much additional effort)*

„The Next Generation is SOLiD“

- Product specification
- SOLiD principles: pdf 15
- Two Base Encoding: pdf 9
- Webinars: SOLiD System Chemistry, Two Base Encoding , see Homepage AB

Assembling of short reads

- Next generation sequencing technology: more data, shorter reads
- Even bigger problems with repeats
- Specific types of sequencing errors

Assembly

Forge Assembler (Platt & Evers)

- New assembler: Forge
- Merge Sanger, 454, Illumina, SOLiD
- PC/Linux cluster
- MPI (Message Passing Interface)

Forge Assembler ... contd.

- „overlap – layout – consensus“
- Overlap graph:
  - nodes = reads
  - edges = overlaps between reads
- 15-mer footprints as seed for 454 data
- exclude overrepresented reads
- linearization of the graph

Metagenomics

- Study of genomes recovered from environmental samples as opposed to from clonal cultures, Wikipedia
- Sargasso Sea Project
  - webinar (C. Venter)
  - press release (NCBI): 1,2 M proteins (ORFs)
- Ultrafast sequencing technologies / Whole Genome Shotgun Sequencing

MEGAN analysis of metagenomic data

- Bioinformatics challenge: methods for identifying taxonomial content of environmental samples
- METaGenome Analyzer, Uni Tübingen, free
- MEGAN paper

MEGAN approach
(LCA approach)

1. Compare reads to databases, such as NCBI-nr or NCBI-nt, using BLAST(N,X,Z).
2. Determine all taxa matched by a read
3. Find the lowest node in the NCBI taxonomy that encompasses all hits of a given read - assign the read to this node (i.e. to this taxon)

Computational chracterization of short environmental DNA fragments

- Identify gene fragments by alignment to Pfam
- strategy to find suitable reads: paper (?)
- „Environmental gene tags“ characterize the sample:
  - Simulation of 454 reads (readsim) from 77 complete genomes
- Conclusion: 454 sequencing can be used to characterize the genetic diversity & taxonomic composition of microbial communities

New Sequencing Strategies for Microbes

- Richardsen, Joint Genome Institute (JGI)
- Most efficient manner to sequence microbial genomes (150 genomes done)
- Hybrid 454/Sanger approach:
  - coverage: 15 x 454 and 4-5 x Sanger
- Goal: eliminate Sanger completely

Strategy of hybrid sequencing

- assemble 454 reads into contigs (Newbler)
- shear (in silico) into 1000 bp overlapping fragments (pseudo-Sanger)
- Assemble together with Sanger reads using Phrap, PGA, Arachne
- Realign with original 454 contigs

Ultrafast *de novo* sequencing of the human pathogen *Corynebacterium urealyticum* with the Genome Sequencer System

Feature	Value
No. of assembled bases	2,294,755
No. of assembled contigs (> 500 bp)	69
Mean G+C content of DNA	64.4%
No. of ribosomal RNAs	5
No. of transfer RNAs	46
No. of protein-coding sequences	2027 (100%)
No. of proteins homologous with proteins of *C. jeikeium*	1589 (78.4%)
No. of proteins non-homologous with proteins of *C. jeikeium*	438 (21.6%)
Coding density	90.2%
Mean size of coding sequences	1036 bp
Mean size of intergenic regions	136.7 bp

Table 1: General features of the *Corynebacterium urealyticum* genome.

paper

Corynebacterium ... *continued*

- Synteny analysis at the protein level with other corynebacterial genomes
- Evidence for genomic rearrangements
- „Bidirectional best BlastP hit analysis“ of genome synteny, see Sybil package
  - identify orthologous genes between genomes
  - groups together any pair of proteins for which each is the other's best BLASTP hit
  - problem: families of closely-related paralogs

Access to the plasmid mobilome of wastewater treatment plant bacteria by applying the 454-sequencing technology

- wastewater treatment plants are a reservoir for bacteria harbouring antibiotic resistence plasmids
- 350.000 454 reads
- 49.000 reads could be mapped to known plasmid genes deposited in the databases
- Annotation with: GenDB, SAMS
- Results: many reads represent genes involved in plasmid replication, mobilisation, and stability

Next Generation Sequencing: Comparison of the technologies for bacterial genome sequencing

- K. Stangier, GATC, Konstanz, Germany
- A bacterial genome (6.5 MB) has been sequenced using 3 technologies: Sanger, Roche/454, Illumina/Solexa
- Recommendations:
  - de-novo assemblies: Sanger combined with 454
  - SNP detection: Solexa (ultradeep coverage)
  - re-sequencing: Illumina/Solexa
- „Newbler assembly can be imported to DNASTAR [distributor: GATC] and assembled with ABI 3730 data“

Sequencing wine grape: Pinot Noir

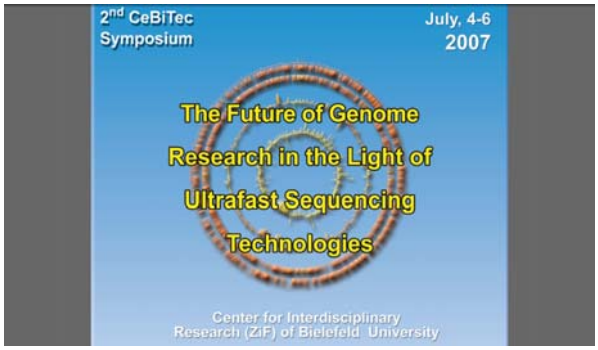
- Methodology:
  - 7 x Sanger; 4.2 x 454
  - „highly parallel“ primer walking
  - metacontigs: end sequences of 2 BAC libraries and a fosmid library
  - 705 metacontigs oriented and ordered using genetic markers (F1 of Syrah x Pinot Noir)
  - Detailed data

Profiling plant transcriptomes by massively-parallel pyrosequencing

- Proof-of-concept study
- Arabidopsis thaliana
- *Step 1*: Two 454 runs on a non-normalized cDNA population provided > 0.5 M ESTs

Profiling plant transcriptomes ... *cont.*

- *Step 2*: Mapping of the ESTs to Arabidopsis genome
  - Transcription of 17.500 genes detected
  - Mapping of ESTs to the corresponding full length transcripts indicate that all regions of the transcripts were represented (regardless of transcript length or expression level)
  - 16.000 ESTs not in dbEST
- However: 26% of all reads were derived from only 25 highly expressed genes ➔ normalisation required (for certain purposes)
- Paper



de novo Sequencing

Newbler™ Assembler is the new *de novo* assembly software developed by 454 Life Sciences™. Exploiting the inherent advantages of the Instrument's performance, the Newbler Assembler operates in flowgram signal space, as opposed to the standard nucleotide space. By operating in flowgram signal space, Newbler Assembler is able to utilize the abundant information stored in the flowgram signals that is lost after base calling (conversion to nucleotide space). Newbler Assembler has three main components: overlap generation, contig layout, and consensus generation. The overlap generator aligns raw reads in flowgram signal space using a proprietary algorithm. Consensus generation is based on **signal averaging** where all aligned flowgram signals at each position are averaged and the final base call is done on the averaged signal. The signal averaging allows higher quality consensus base calls.

Assembly and Annotation of short read data sets

- *Himmelbauer*, MPI for Molecular Genetics
- de-novo assembling of genomic sequences from short-read data
- **SHARCGS** (SHort-read Assembler based on Robust Contig-extension for Genome Sequencing)
- 25-50mer reads
- BAC inserts, yeast chromosomes, bacteria