Applied bioinformatics in genomics

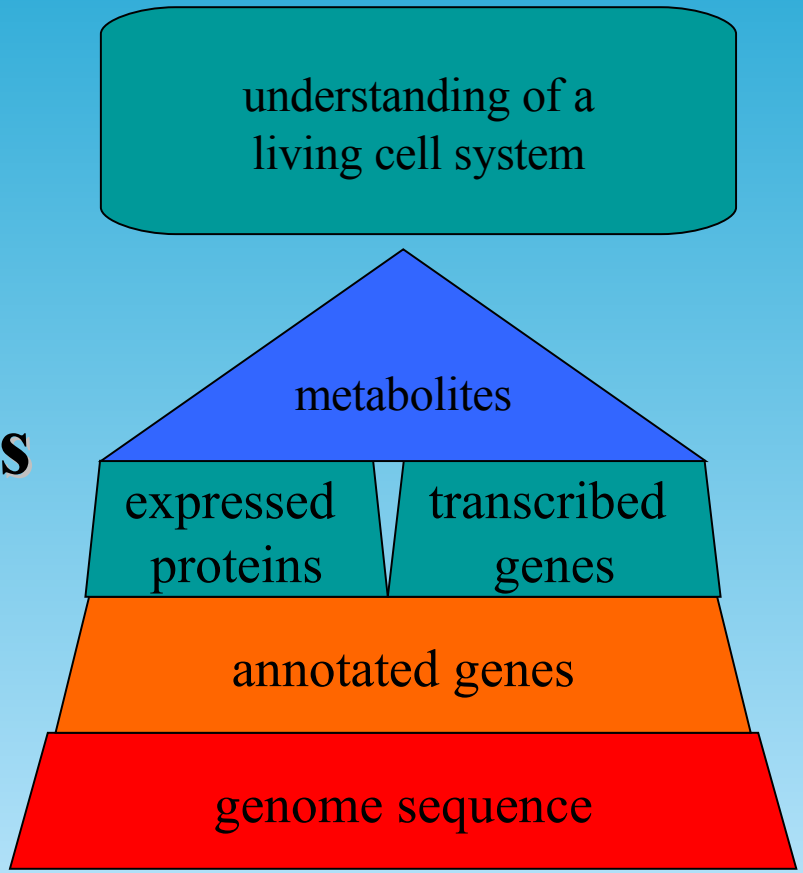
Productive bioinformatics in a genome sequencing center

Heiko Liesegang

Warschau 2005

The –omics pyramid:

١. **Genome sequencing**
٢. **Genome annotation**
٣. **Transcriptomics**
٤. **Proteomics**
٥. **Metabolomics**

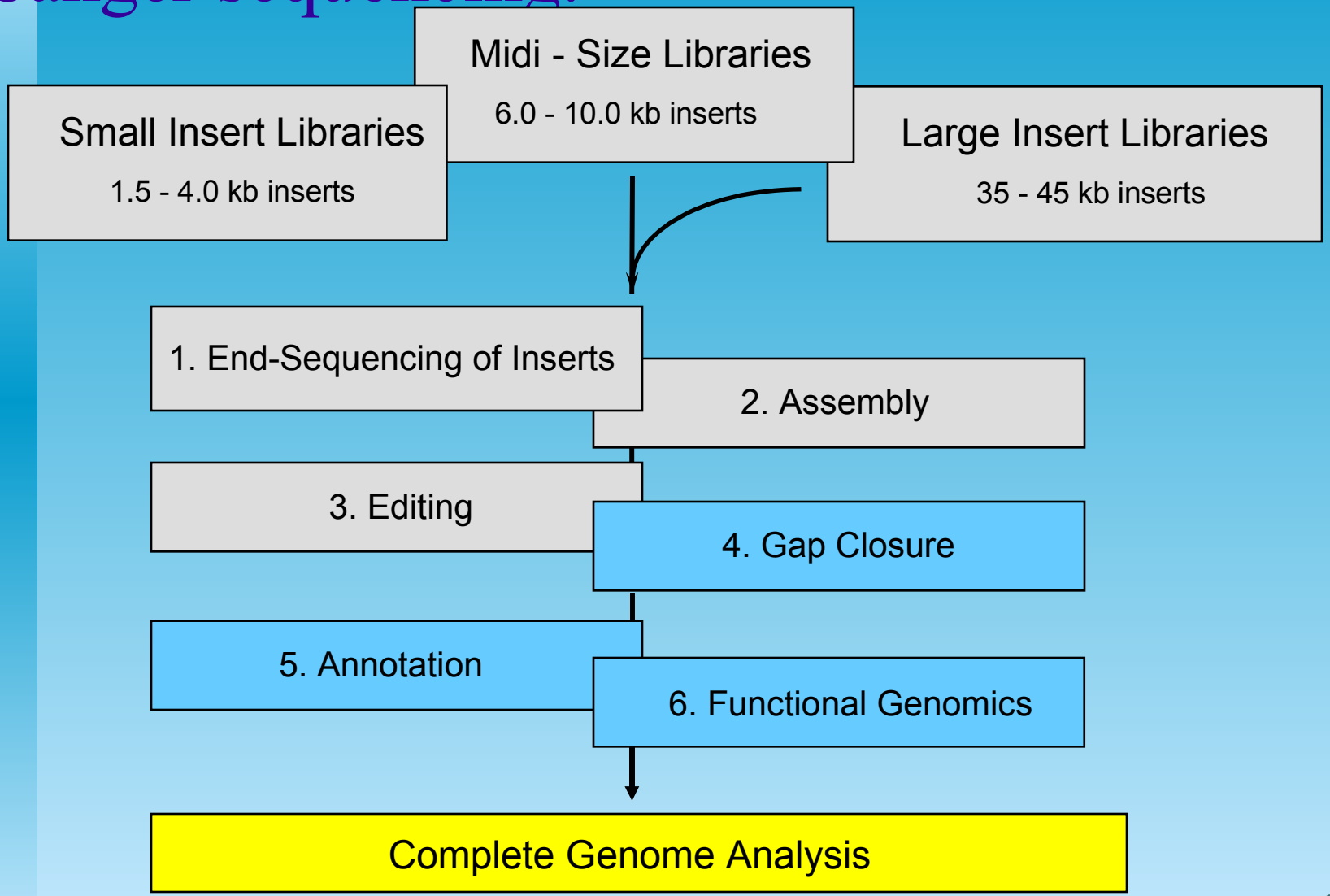


Genome sequencing

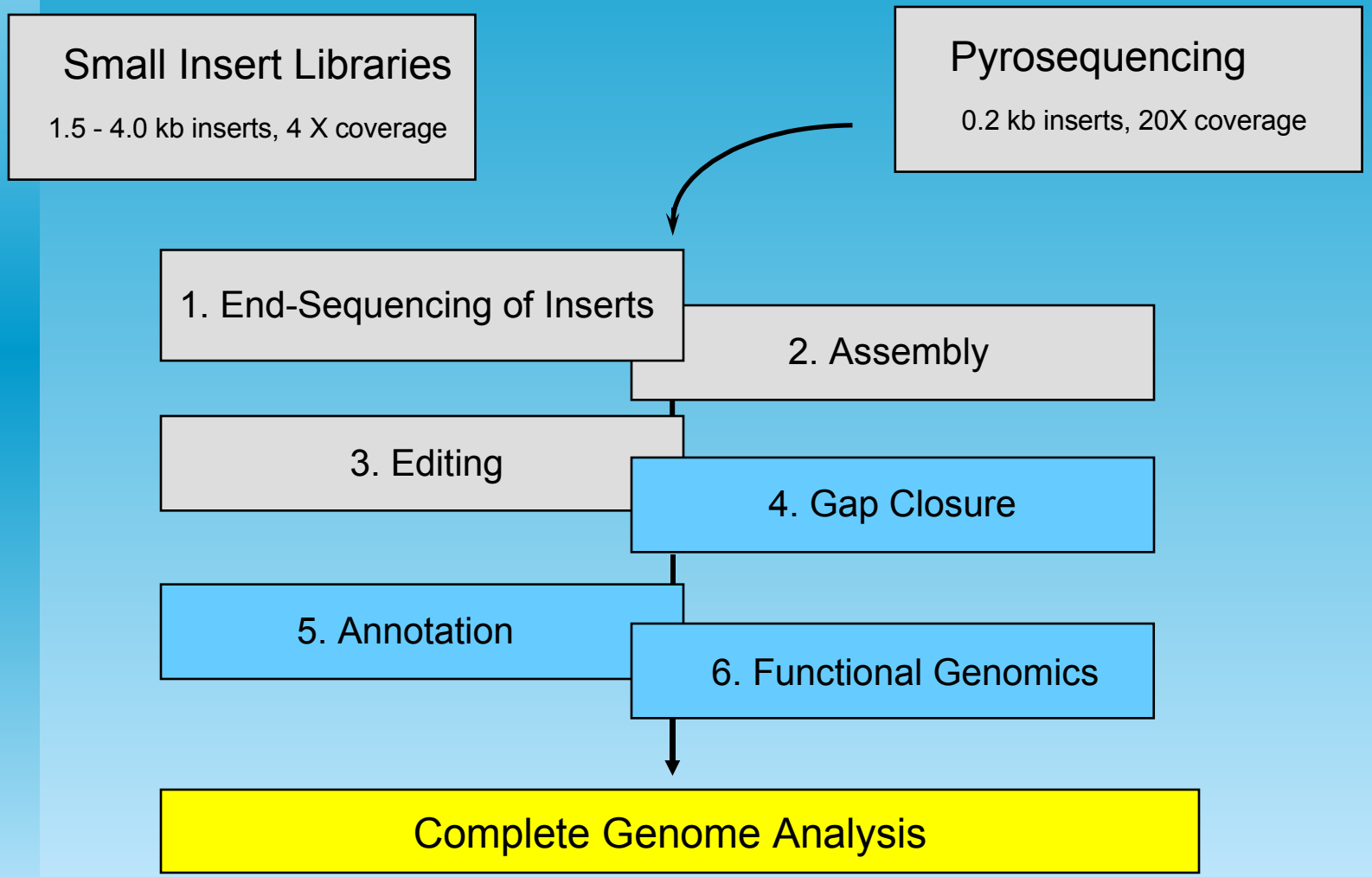
Sequencing Capacities at the G2L

Unit(s)	sequences / week	average read lengths	
1x 454Flex (pyrosequencing)	0,5 x 200.000	~200 bp (low quality, short reads)	60 Mbp
2x ABI 3730XL (96 cap.)	64 x 96	1000 bp (high quality long reads)	6,14 Mbp
	≈ week	≈ Σ 6 Mbp – 66 Mbp	

Genome-Sequencing-Strategy based on Sanger sequencing:

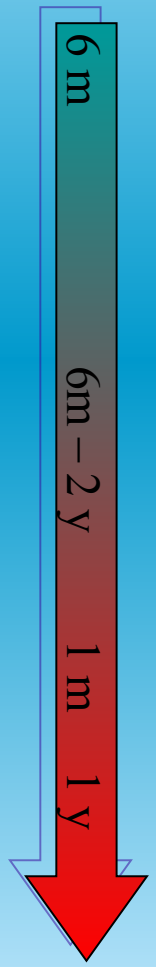


Genome-Sequencing based on mixed Pyro- and Sanger sequencing



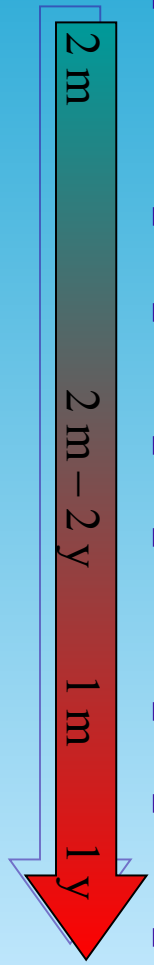
Principle steps in a genome project

Sanger



- **Raw sequencing** – producing random distributed sequences of 8 to 10 fold coverage
- **Processing** – checking the quality of the raw data
- **Assembly** – assemble overlapping reads to reconstruct the replicons of the genome
- **GAP-Closure** – using PCR techniques based on genomic information to generate missing sequences
- **Editing** – manual correction of the computer based assembly
- **Gene prediction** – ORF finding and functional annotation of the DNA-sequence.
- **Annotation** – assign functions to genes to produce biological data

Sanger + 454



- **Raw sequencing A** – producing random distributed sequences of 4 to 5 fold coverage Sanger sequence plus high coverage 454 data ~20 fold coverage
- **Processing** – checking the quality of the sanger raw data
- **Assembly I** – co assemble overlapping reads of sanger sequence with pre-assembled 454 data
- **Assembly II** – reconstruct the replicons of the genome
- **GAP-Closure** – using PCR techniques based on genomic information to generate missing sequences
- **Editing** – manual correction of the computer based assembly
- **Gene prediction** – ORF finding and functional annotation of the DNA-sequence.
- **Annotation** – assign functions to genes to produce biological data

GenoMik Network Göttingen

Bioinformatics: Hardware aspects

Requirements of genome assembly:

8 MBp genome assembly with appr. 80,000 reads sanger sequence takes 12 GByte RAM and appr. 2 h calculation time on 1GHz RISC chip

Requirement of genome annotation:

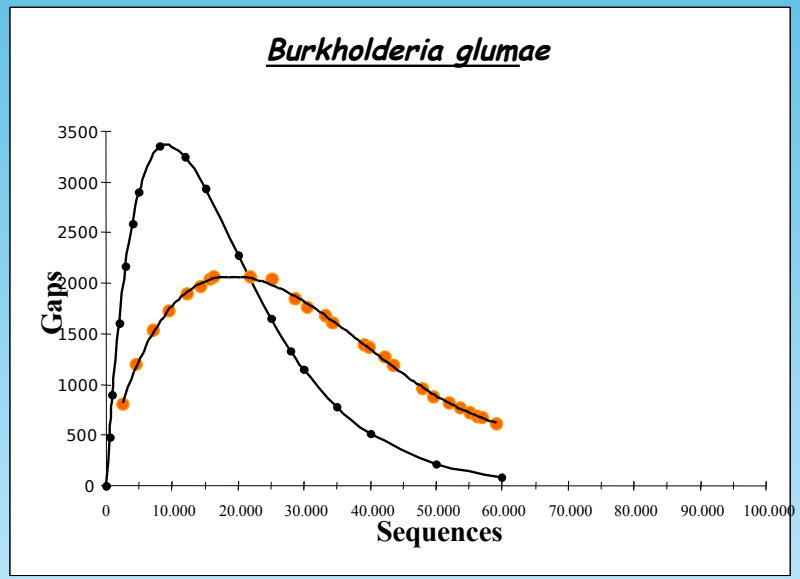
automated similarity based annotation of 1 Mbp genome sequence (~1,000 genes) needs 10 h calculation times on 16 parallel working processors.

Manual annotation:

One human experts annotates 50 genes/d on a web based annotation server

Rawsequencing – the first dataprocessing

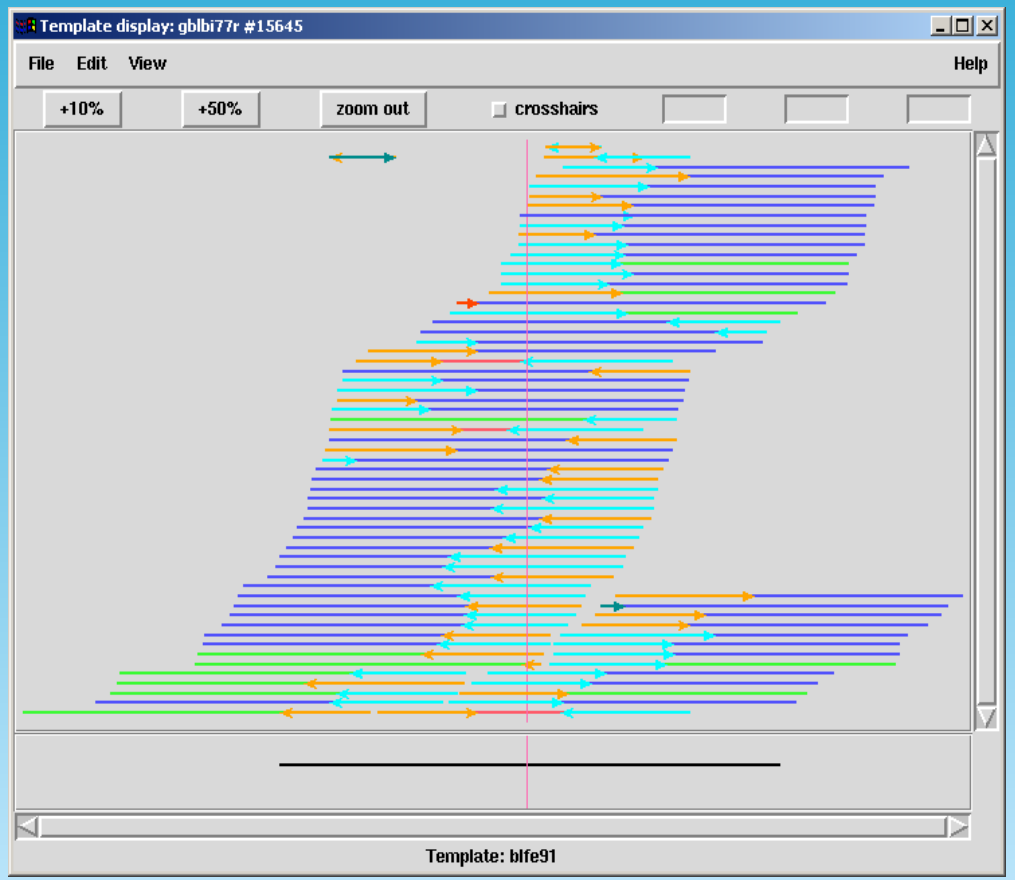
- Check for read quality
- Check gene libraries
- Check genome coverage
- Lander-Waterman plot



Pregap – Screening for contaminating sequences

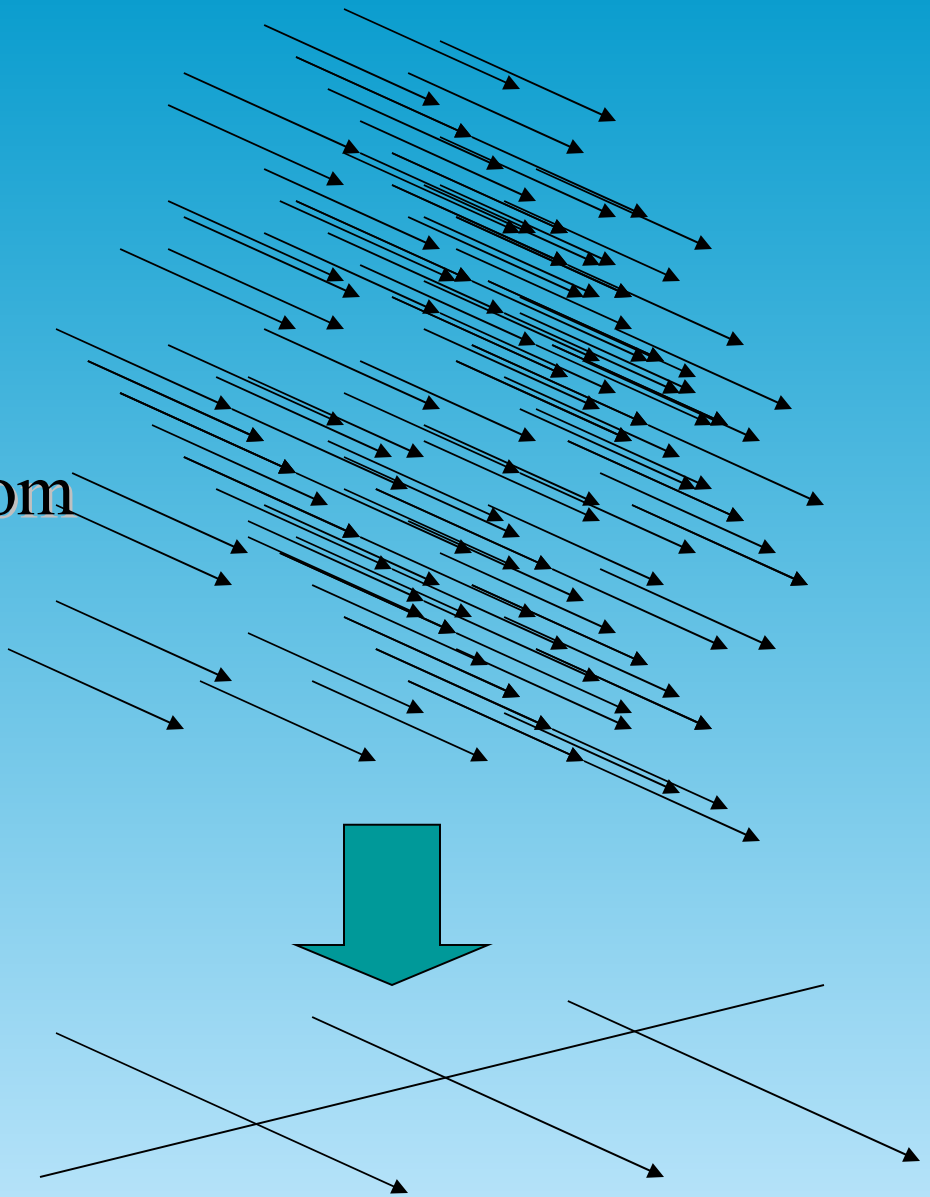
- **Qclip** – defines the minimum quality for segments of sequences used in assembly.
- **Svec** - localises the insertion site at the left end of the sequence.
- **Screenveq** – looks for contaminating vector sequences.

Screenvec failure:
sequencing the vector due to gene library contamination



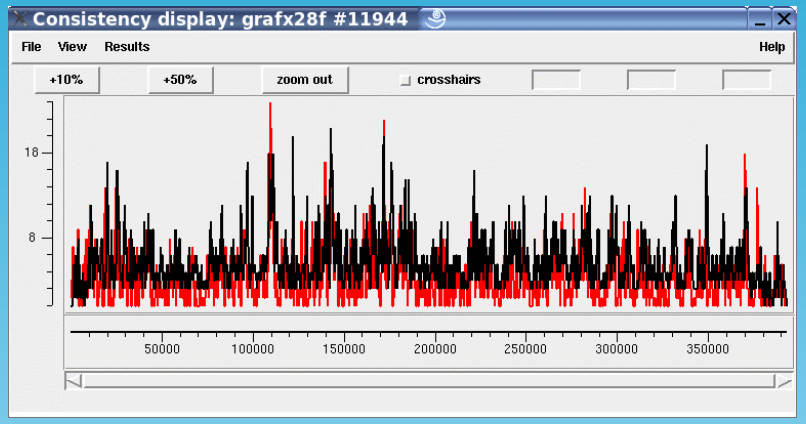
Assembly

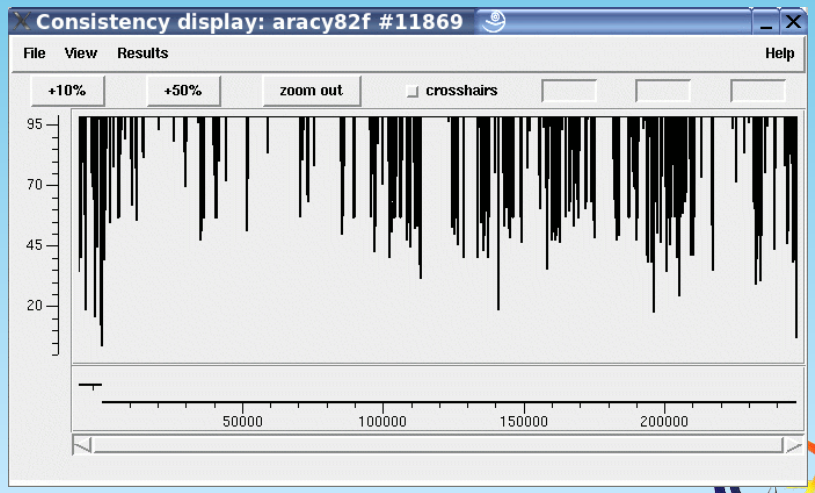
- Gcphrap
- Producing contigs from overlapping reads
- Typically
  10.000 reads / 1Mbp



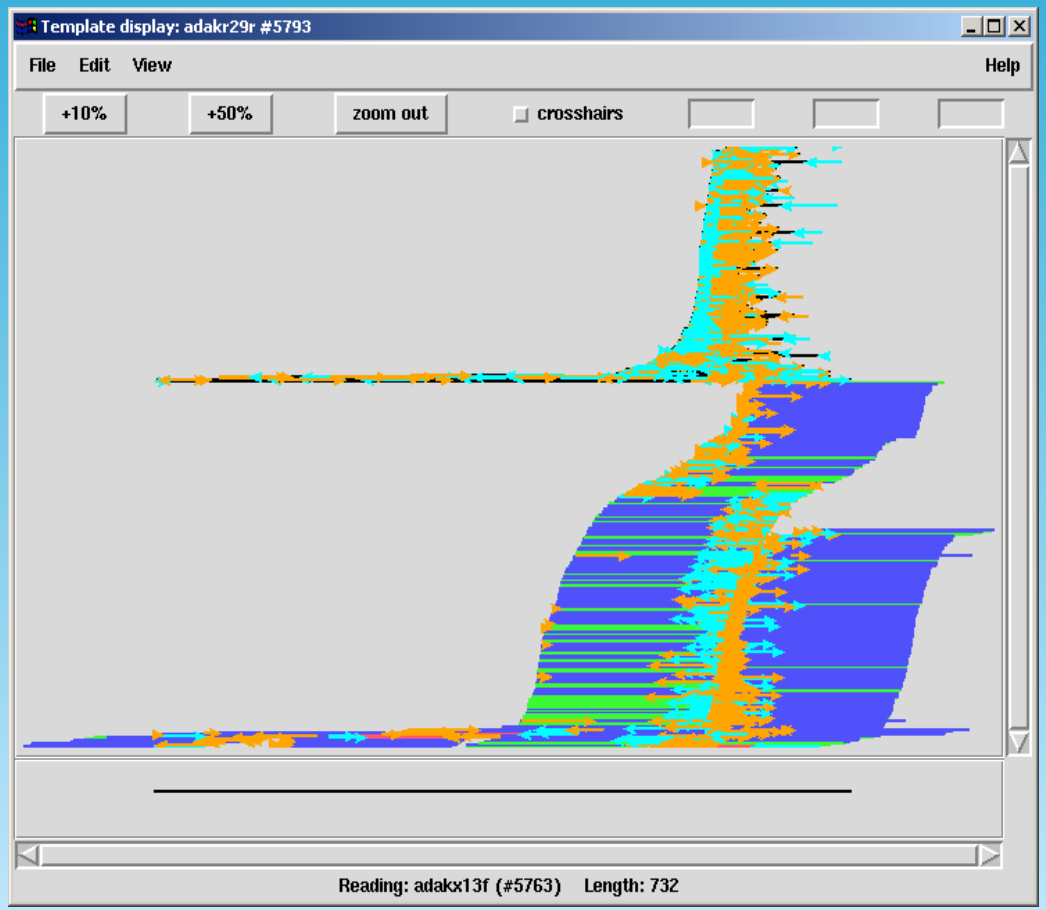
Checking the initial assembly

- Distribution of templates
- Size of templates
- Confidence of the contigs

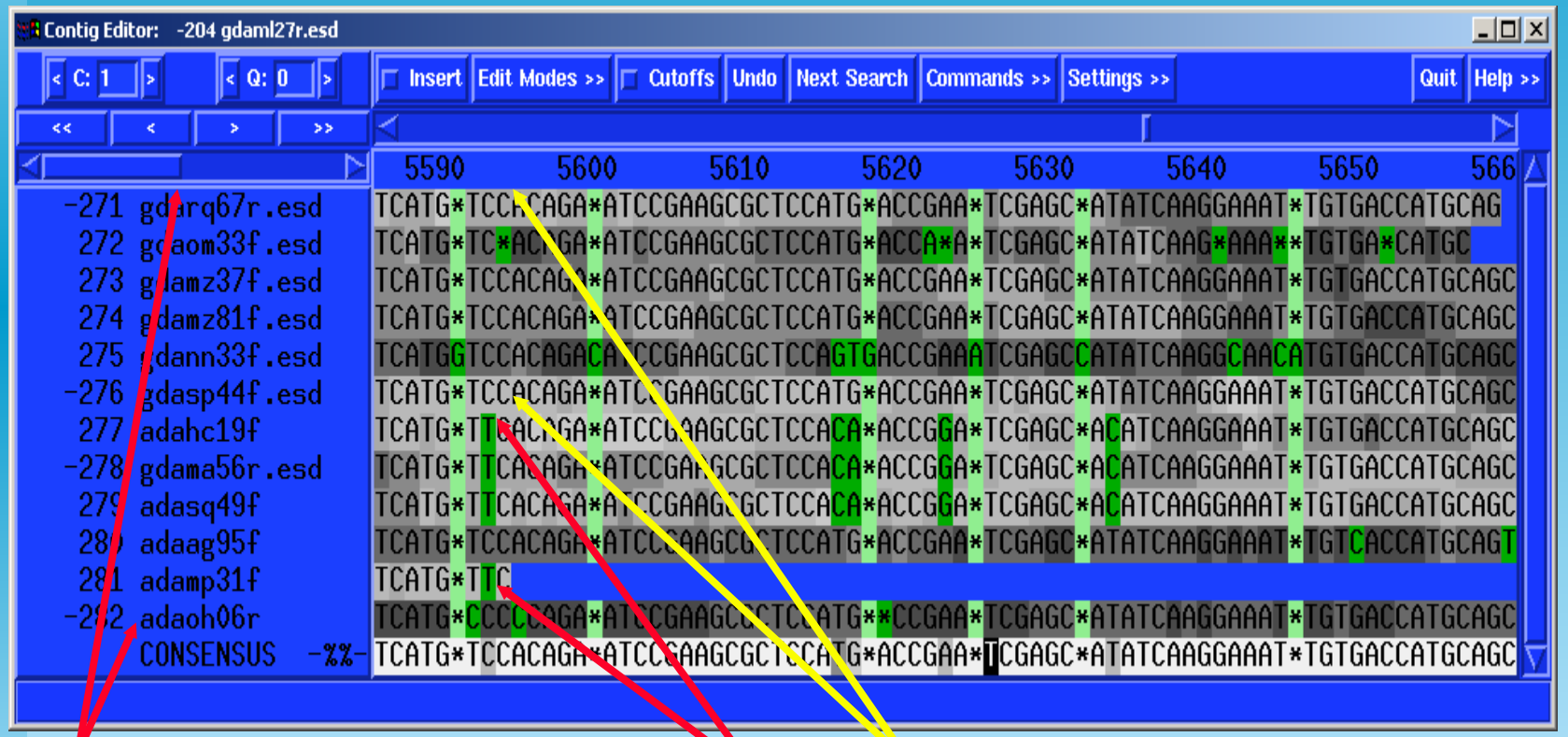




Screenvec failure: Short vector contaminations of ~0.1 –0.9 kbp



Missassembly

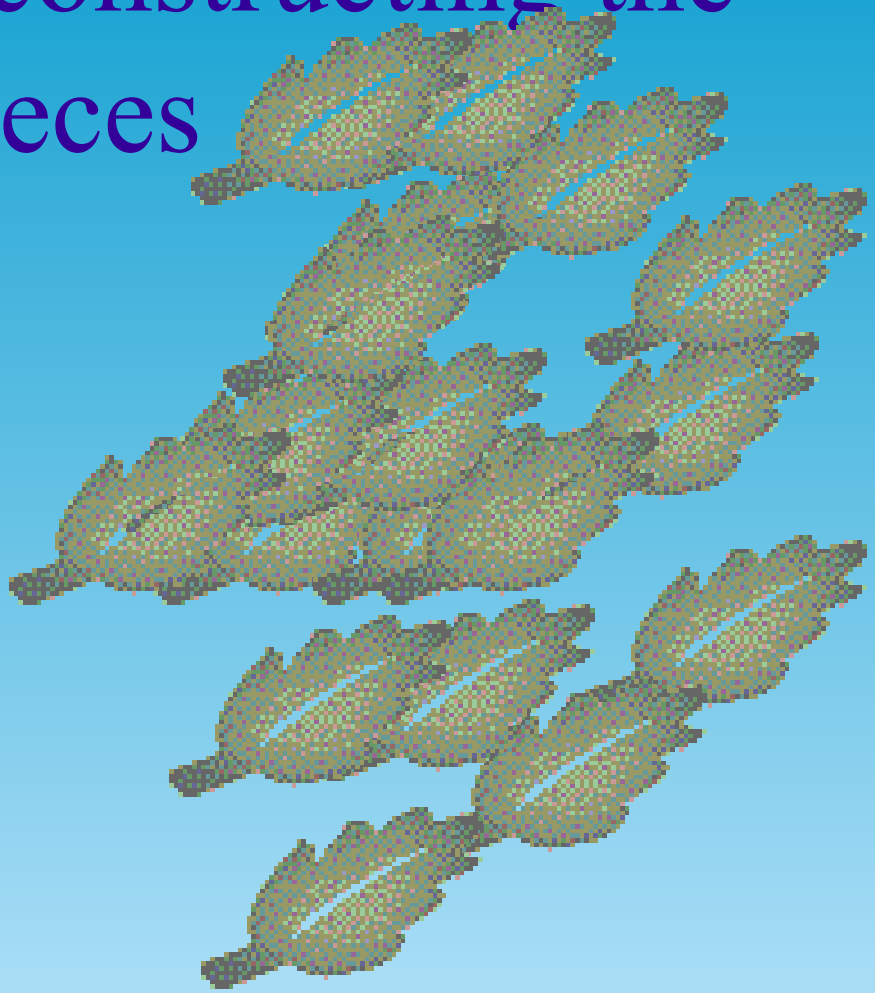


doubled average base coverage

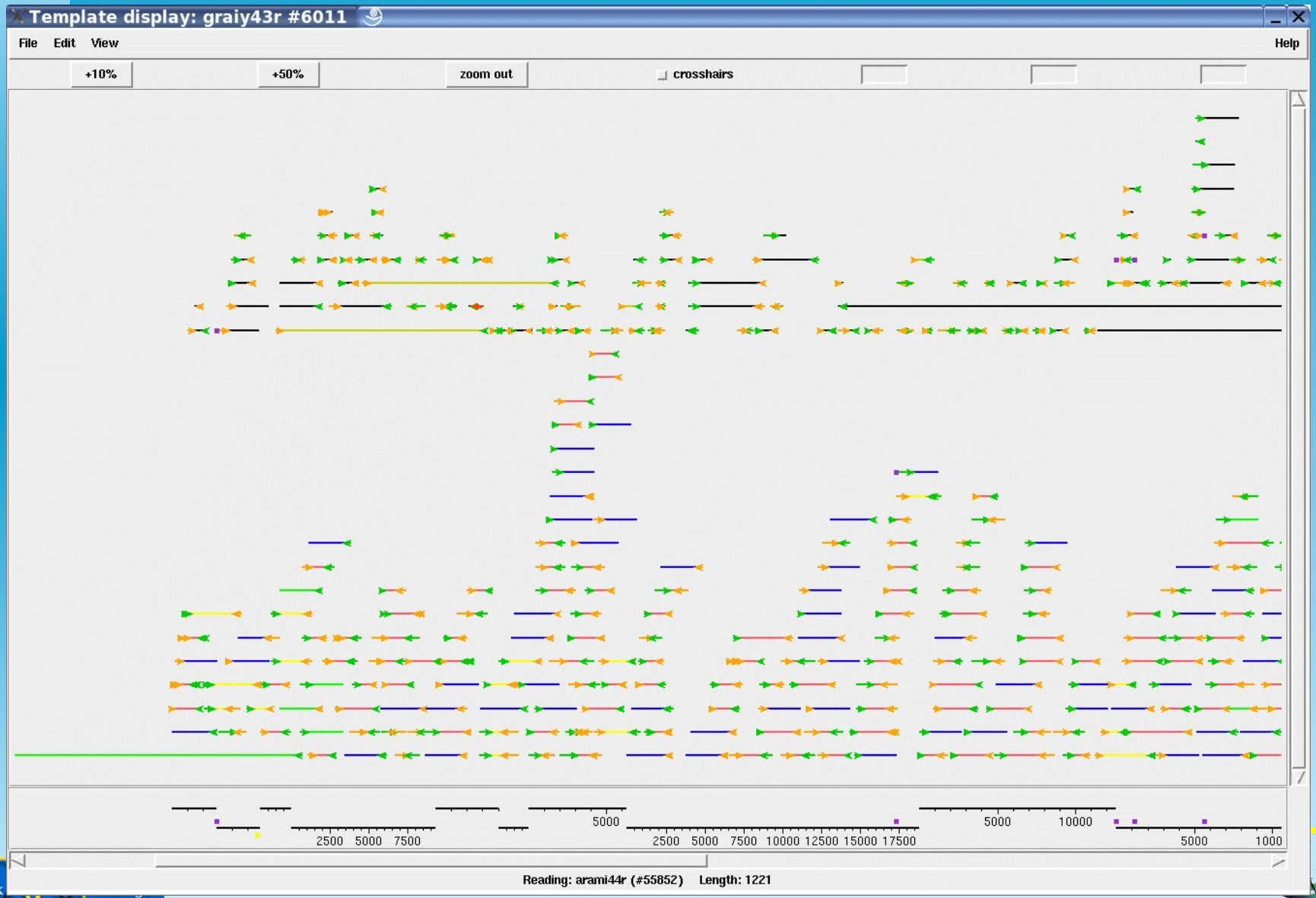
Apparently two groups of sequences in the multi alignment

GAP-closure reconstructing the genome from pieces

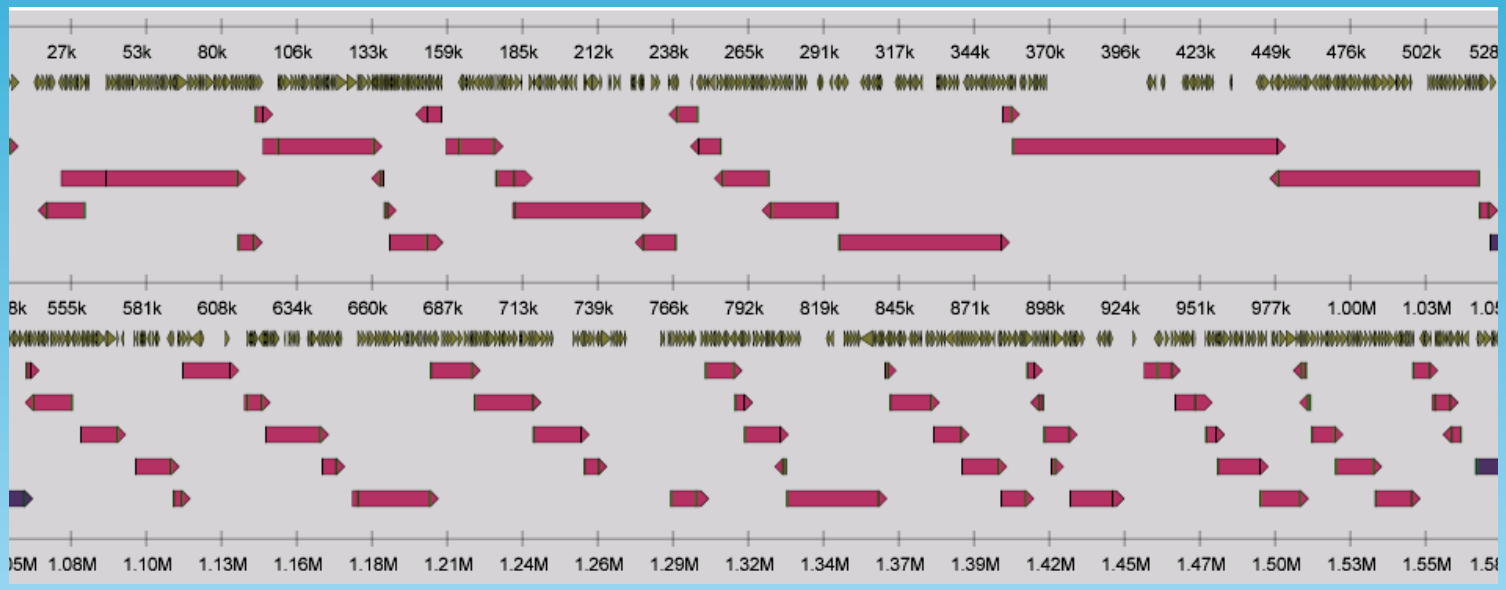
- Sorting contigs
  - Gap spanning templates
  - Projector
  - ERGO gene cluster
- Closing sorted contig gaps by PCR

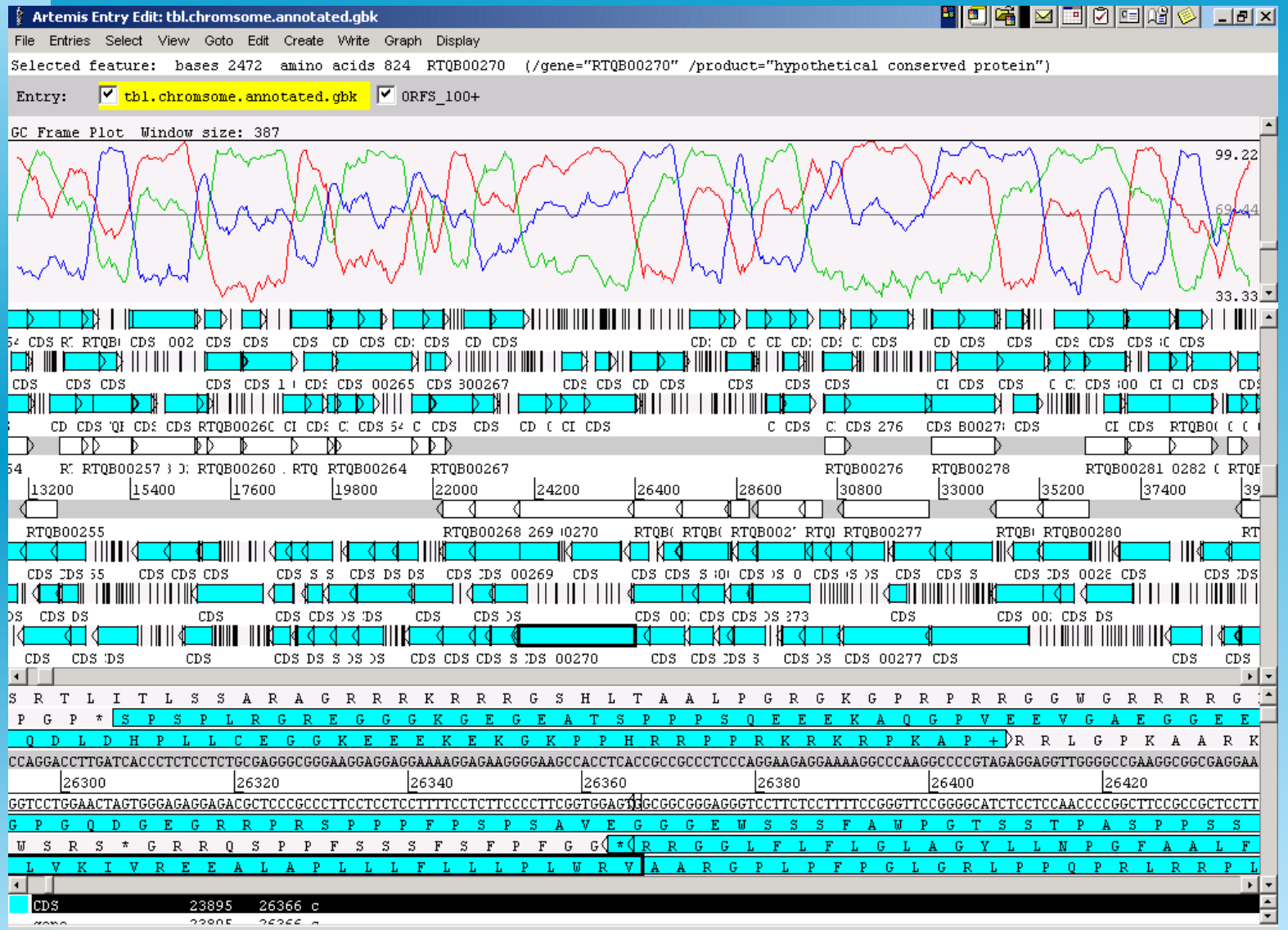


Sorting contigs by templates



Projector contig projection on related genomes

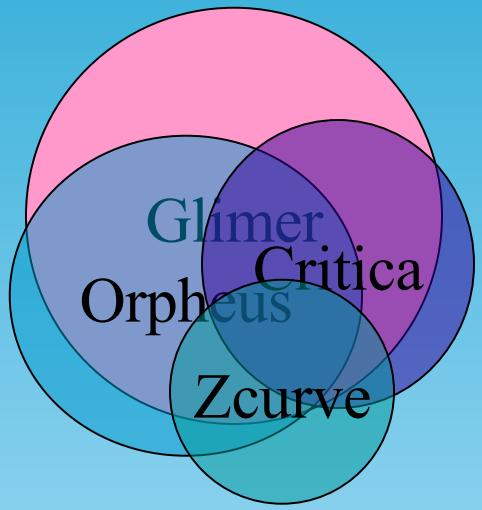



Artemis Entry Edit: tbl.chromsome.annotated.gbk
File Entries Select View Goto Edit Create Write Graph Display
Selected feature: bases 2472 amino acids 824 RTQB00270 (/gene="RTQB00270" /product="hypothetical conserved protein")
Entry: tbl.chromsome.annotated.gbk ORFS_100+
GC Frame Plot Window size: 387
99.22
33.33
13200
15400
17600
19800
22000
24200
26400
28600
30800
33000
35200
37400
RTQB00255
RTQB00268
RTQB00269
RTQB00270
RTQB00277
RTQB00280
26300
26320
26340
26360
26380
26400
26420
CDS 23895 26366 c

YACOP - Combined ORF finding

- Glimmer
- Orpheus
- Critica
- Zcurve

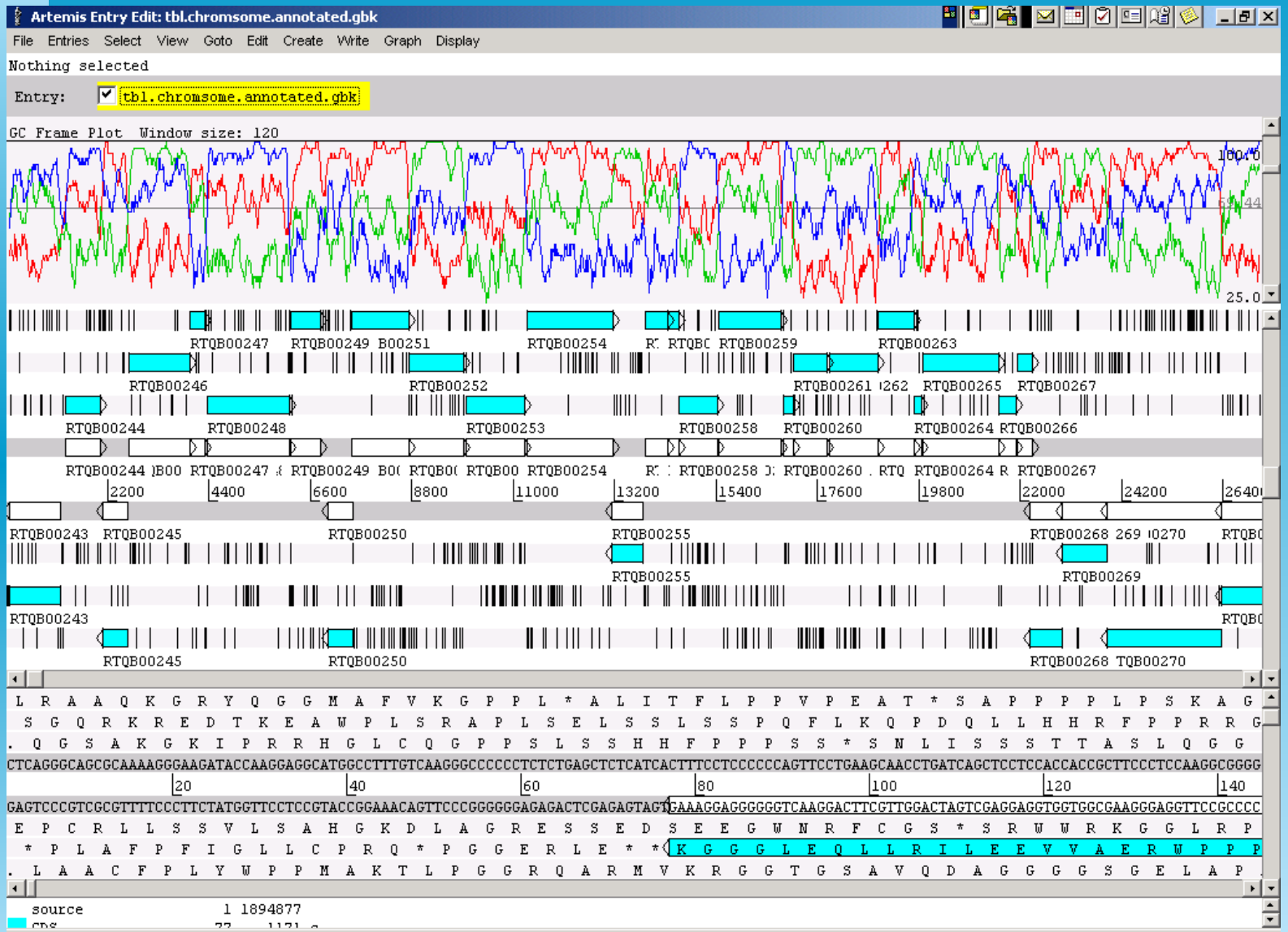
YACOP



- ORF finding is based upon statistical properties of coding regions
- Experts have to curate ORFs according to frameshifts, startcodons, artifact ORFfinding

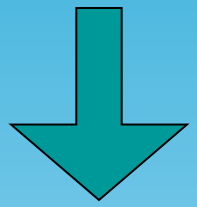
ORF to gene

- *Thermus thermophilus* sequences bare 2 – 3 fold coverage of ORFs per sequence
- Automated ORF finding uses codon biases und HMM to identify potential “coding” ORFs
- Gene start prediction is curated by RBS-finder
- Accuracy: ca. 97% relevant ORFs can be identified ca. 50% to 65% of gene starts can be correctly predicted
- All ORFs have to be curated by and human experts.


Artemis Entry Edit: tbl.chromsome.annotated.gbk
File Entries Select View Goto Edit Create Write Graph Display
Nothing selected
Entry: tbl.chromsome.annotated.gbk
GC Frame Plot Window size: 120
100.0
64.44
25.0
RTQB00247 RTQB00249 B00251 RTQB00254 RTQB00259 RTQB00263
RTQB00246 RTQB00252 RTQB00261 RTQB00265 RTQB00267
RTQB00244 RTQB00248 RTQB00253 RTQB00258 RTQB00260 RTQB00264 RTQB00266
2200 4400 6600 8800 11000 13200 15400 17600 19800 22000 24200
RTQB00243 RTQB00245 RTQB00250 RTQB00255 RTQB00268 RTQB00270
RTQB00255 RTQB00269
RTQB00243
RTQB00245 RTQB00250 RTQB00268 TQB00270
L R A A Q K G R Y Q G G M A F V K G P P L * A L I T F L P P V P E A T * S A P P P P P L P S K A G
S G Q R K R E D T K E A W P L S R A P L S E L S S L S S P Q F L K Q P D Q L L H H R F P P R R G
. Q G S A K G K I P R R H G L C Q G P P S L S S H H F P P P S S * S N L I S S S T T A S L Q G G
CTCAGGGCAGCGCAAAAGGGAAGATACCAAGGAGGCATGGCCTTTGTCAAGGGCCCCCCTCTCTGAGCTCTCATCACTTTCCTCCCCCCAGTTCCTGAAGCAACCTGATCAGCTCCTCCACCACCGCTTCCCTCCAAGGCGGGG
20 40 60 80 100 120 140
GAGTCCCGTCGCGTTTTCCCTTCTATGGTTCCTCCGTACCGGAAACAGTTCCCGGGGGGAGAGACTCGAGAGTAGTGAAAGGAGGGGGGTCAAGGACTTCGTTGGACTAGTCGAGGAGGTGGTGGCGAAGGGAGGTTCCGCCCC
E P C R L L S S V L S A H G K D L A G R E S S E D S E E G W N R F C G S * S R W W R K G G L R P
* P L A F P F I G L L C P R Q * P G G E R L E * * K G G G L E Q L L R I L E E V V A E R W P P P
. L A A C F P L Y W P P M A K T L P G G R Q A R M V K R G G T G S A V Q D A G G G G S G E L A P
source 1 1894877

T. thermophilus

- T. thermophilus contains two replicons :
  - Chromosome (1,894,877 bp, 69.2% GC content)
  - pTT27(232,605 bp, 69.4% GC content)

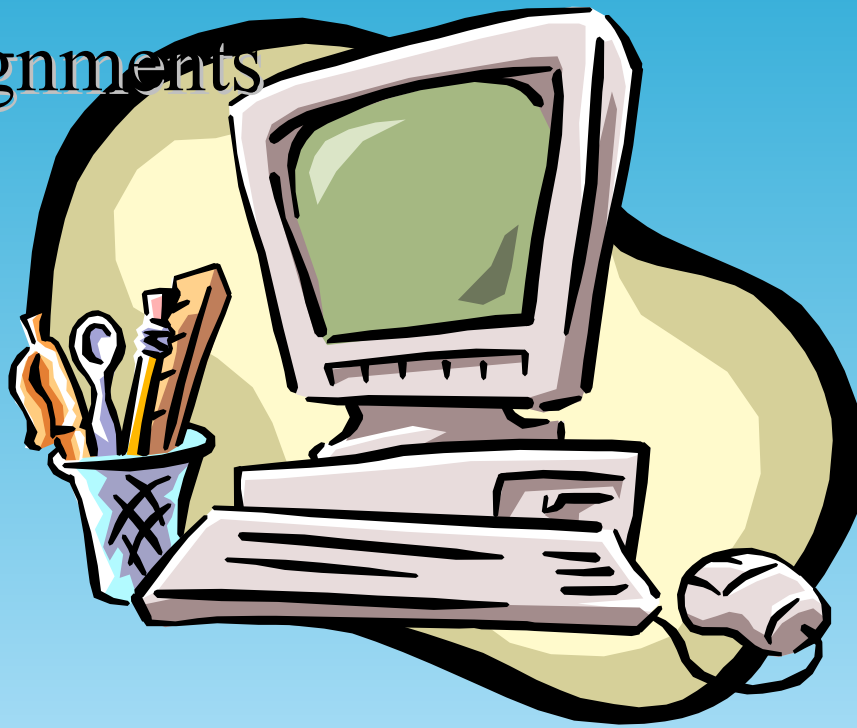


Orffinding,
manual curation

ORFs	Chromosome	pTT27
with assigned function	1397 (70%)	85 (37%)
conserved hypothetical	192 (10%)	56 (24%)
no database match	399 (20%)	89 (39%)
Total	1988	230
rRNA – Cluster	2	-
tRNA	47	-
IS-Elements	30	23

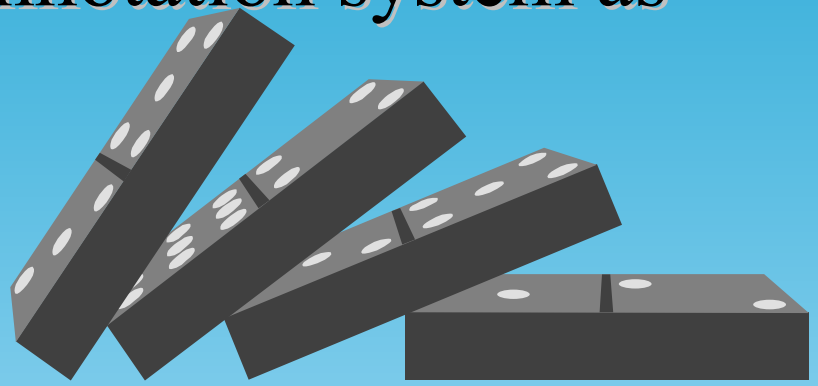
The first Annotation steps- similarity based automated assignments

- FASTA - automated assignments
  - for all proteins against a high quality dataset SwissProt, PIR
- BLAST, PSI-Blast
  - Providing data for manual control



Transitive error propagation

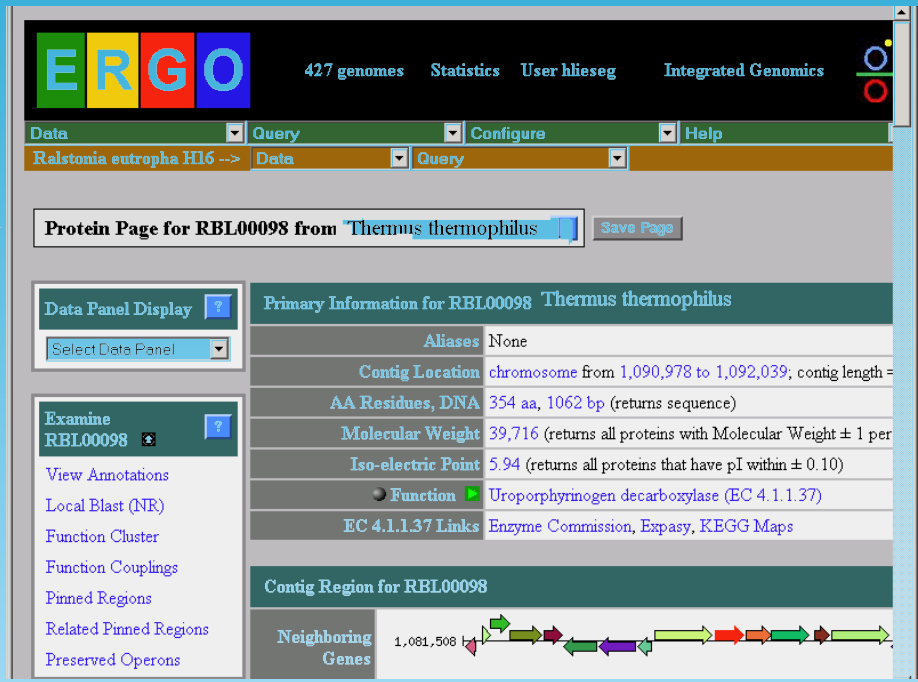
- In automated annotation system as
- Genequiz
- PEDANT
- ERGO/WIT



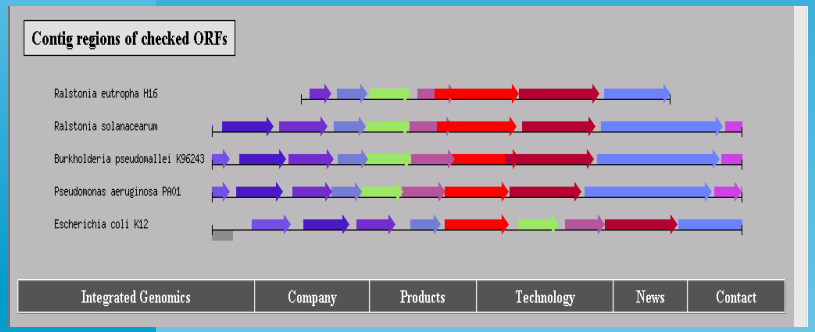
- Annotations are assigned if proteins share a cut of similarity
- => original errors thus multiply by each new entry

Manual Curation of automated Annotations

- A team of annotation experts has to check each automated annotation
- ERGO provides additional non similarity based means to evaluate Annotations
- Tmpred, Pfam, COGs, Prosite are used to achieve a consistent functional assignment
- The kind of assignment evidence is a way of classifying annotation reliability

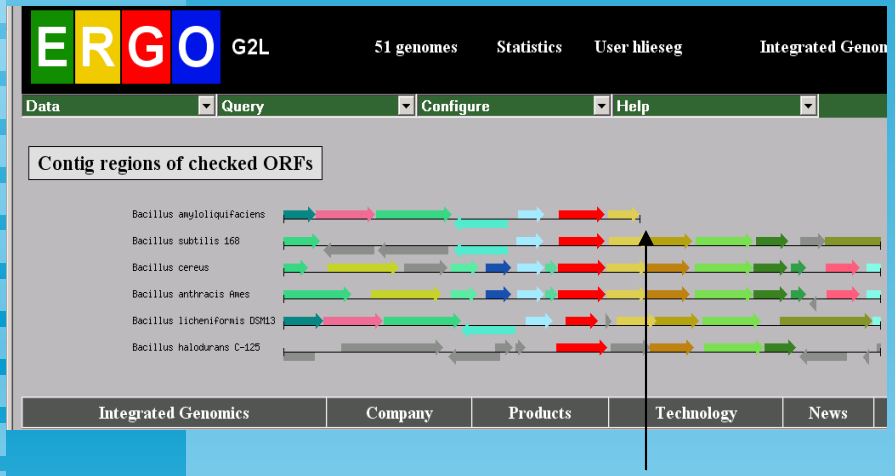


Functional coupling in genomes

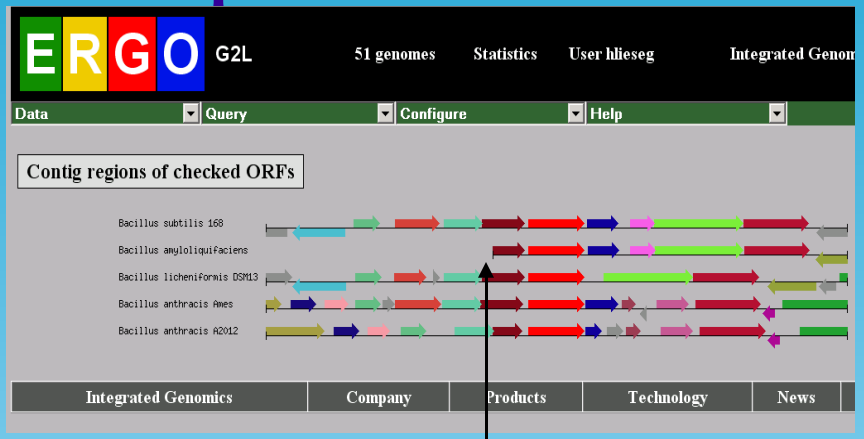


- Most bacterial genes are clustered.
- Related organisms share the organization of gene clusters
- Cluster provide a context of genes for reannotation
- Cluster provide a datasource for gapclosure

Alignment of clustered genes from distant related species



End of one contig inside a gene cluster



End of another contig inside the same gene cluster

Metabolic reconstruction

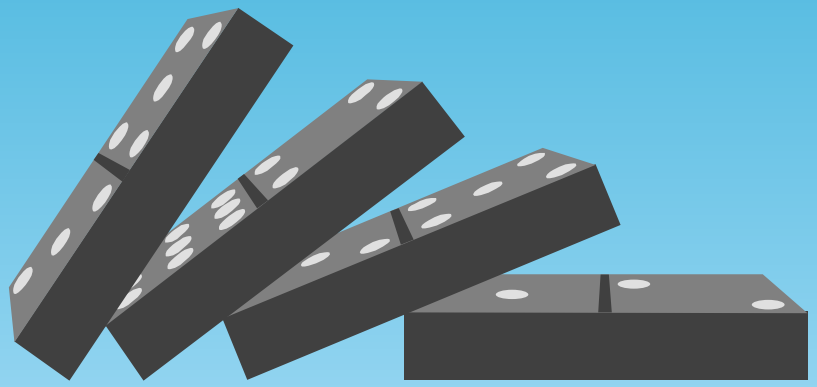
- Based on Annotated genes metabolic pathways can be predicted
- Pathway finding is based upon EC categorisation
- Missing EC functions have to be searched in the “hypothetical” genes
- Thus pathway reconstruction leads to a reannotated genome

Genomic properties

- Genome sequences provide physiological predictions
- Predicted functions have to be proven experimentally
- High throughput evidence methodes from transcriptomics and proteomics give exists/not exists data
- Functional evidence have to be provided by classical wet lab experiments

Source of annotation errors

- ORF-finding errors
- Sequencing errors
- Artifact similarity
- annotation errors in reference database
  - Misspelling
  - Underannotation
  - Overannotation
  - Missannotation



Error in published genomes: Enzymes an EC classification

- The EC classification for enzymes characterizes Enzyme by a four digit number
- EC 6.1.1.19 (=) Arginyl-tRNA Synthetase
  (first characterized as Arginyl-tRNA Transferase, a complete different reaction type)
- **(enzymetype).(cofactors).(reactiontype).(substrate)**
- Errors found:
- First digit class of enzyme (~8 % errors)
- Last digit substrate specifity (>= 33 % errors)

Scaling of annotation errors: Transitive Annotation-Based Scale

Description	Score	description
False positive	7	wrong function
Over-prediction	6	over detailed annotation
Domain error	5	domain error
False negative	4	function not found
Under prediction	3	under detailed annotation
Undefined	2	-
Typographical error	1	misspelling

Chlamydia trachomatis in different annotations

Score	Original	Genequiz
7	43	35
6	50	62
5	20	23
4	26	50
3	67	48
2	84	64
1	35	10
0	565	598

Ranking of reliability

1. Biochemical experimental characterization
2. Genetic experimental data from mutants
3. Consistent properties to a highly similar database hit
4. Good database hit
5. Consistent frame according to the used ORFfinder
6. Nonoverlapping ORF of minimum size

YACOP – gene prediction pipeline in G2L

- Gene prediction is based upon different aspects of typical genes.
- YACOP uses an combined ORF-finding of Glimmer, Zcurve and Critica with a startcodon correction by RBS-finder
- All ORFs a curated manually by human experts.
- In genetic islands genes have still to be predicted by human experts, automated ORF-finding fails almost complete.

Gene prediction – looking for biological features

Protein coding ORFs ca. 90 – 99 % (only 65% correct starts)	YACOP, TICO, GeneMark, Glimmer, ...
tRNA ca. 99%	tRNA-Scan
rRNA ca. 99%	BLASTN
ncRNA ?	structure based evaluation of short alignments

Annotation

- Assigning functions to sequences
- **Proteins:**
- GenDB
- ERGO
- **Islands/prophages:**
- SIGI/Colombo
- **Transcriptomics:**
- Which genes are working?