

DNA SEQUENCING

- the past, the present and the future

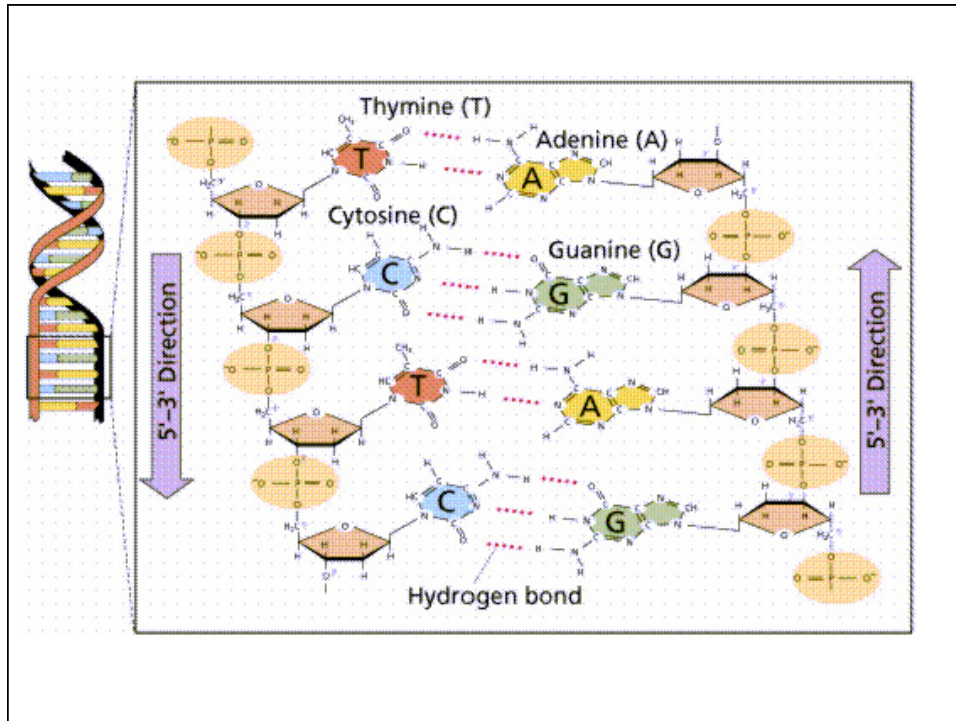
Anders Blomberg
Dept. of Chemistry and Molecular Biology
anders.blomberg@cmb.gu.se



GÖTEBORGS UNIVERSITET

Historical note: Proof that DNA is the Genetic Material

- Griffith - **1928**
When he added dead, virulent pneumonia bacteria to a mouse, it lived; but if he first added dead virulent bacteria to live non-virulent bacteria and then to mice, some mice died. He termed the material that changed the non-virulent bacteria to virulent **the transforming principle**.
- Avery, MacLeod and McCarty - **1944**
They used biochemical purification of cellular fractions to determine that DNA and not RNA or protein was the transforming principle.
- Watson and Crick - **1953**
The structure of DNA; double helix
The structures implication for chromosome replication (duplication)



FREDERICK SANGER

Since 1940 he has carried out research in the Department of Biochemistry at Cambridge. From 1940 to 1943 he worked with Dr. A. Neuberger on the metabolism of the amino acid lysine and obtained a Ph.D. degree in 1943.



The Nobel Prize in Chemistry 1958

"for his work on the structure of proteins, especially that of insulin"



The Nobel Prize in Chemistry 1980

"for his contributions concerning the determination of base sequences in nucleic acids"



SANGER'S DIDEOXY METHOD - 1975

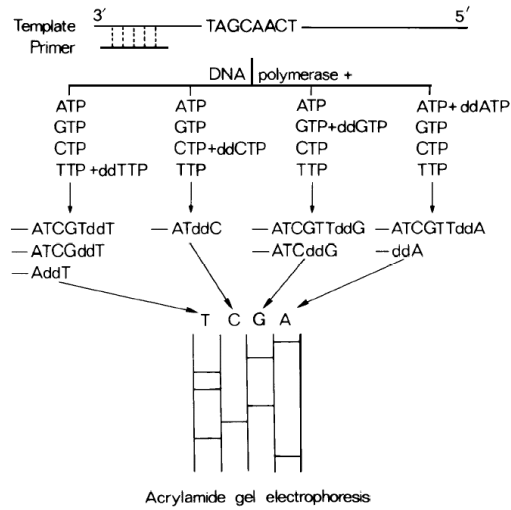
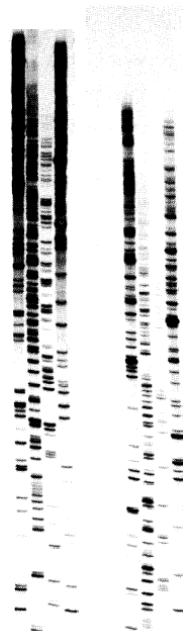


Fig. 3. Principle of the chain-terminating method.

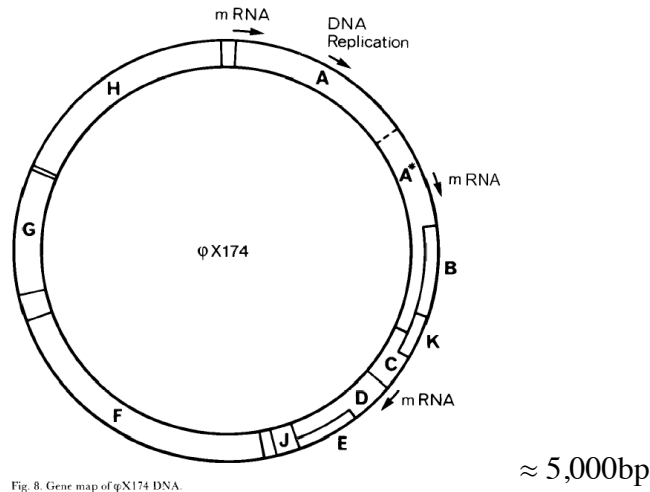
25hr
GATC 5hr
GATC



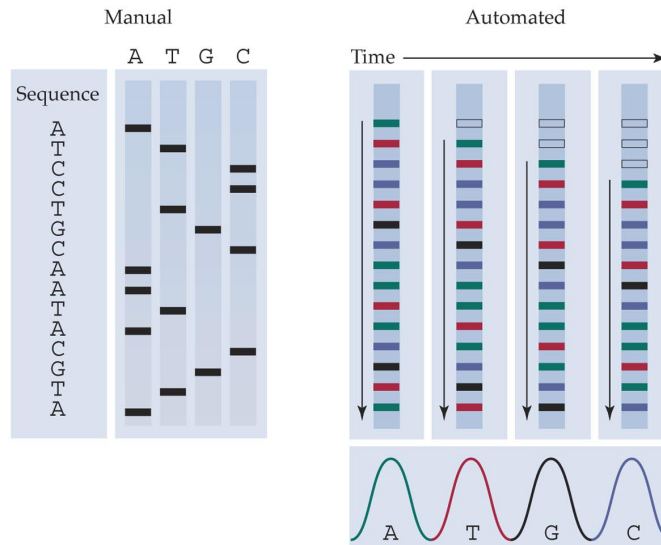
SEPARATION OF
FRAGMENTS BY
ELECTROPHORESIS
- single nucleotide resolution

THE FIRST SEQUENCED GENOMES – 1977

Virus



Automation – faster and faster and....



A PRIMER OF GENOME SCIENCE, Second Edition, Figure 2.1 (Part 2) © 2005 Sinauer Associates, Inc.



CAPILLARY ELECTROPHORESIS SEQUENCING

- Average read 500 - 900 bp
- Many capillaries - up to 384
- > 2Mbases per day/machine
- 96, 384 well plates



SEQUENCING CENTRES





Cost of DNA sequencing

- | | | |
|------------------------|--------|--------|
| • 100SEK per base | (1990) | Sanger |
| • 1SEK per base | (2004) | |
| • 0.1SEK per base | (2006) | |
| • 0.0001SEK per base | (2008) | NGS |
| • 0.00001SEK per base | (2010) | |
| • 0.000001SEK per base | (2011) | |

\$ 1,000 human genome

J. Craig Venter Science Foundation Announces \$500,000 Technology Prize for Advances Leading to the \$1,000 Human Genome

ROCKVILLE, MD (September 23, 2003). The J. Craig Venter Science Foundation announced today a \$500,000 Genomic Technology Prize. The prize, to be awarded one time only, is aimed at stimulating the scientific and technology research community to significantly advance automated DNA sequencing so that a human genome can be sequenced for \$1,000 or less as soon as possible. The prize was announced during New Frontiers in Sequencing Technology session at the 15th annual Genome Sequencing and Analysis Conference (GSAC) in Savannah, Georgia.

Over the last decade there have been significant advances in the field of genomics. More than 150 genomes, including the human genome, have been sequenced. Despite this progress we need substantial improvement in technology so that genomics can be fully integrated into all of our lives. One such area is DNA sequencing, said J. Craig Venter, Ph.D., president and founder of The J. Craig Venter Science Foundation. By continuing to reduce the cost and increase the accuracy and speed of DNA sequencing we will enable genomics to be more fully integrated into areas such as clinical medicine. It is the hope of the Venter Science Foundation that providing this challenge to the scientific community will enable us to reach the \$1,000 genome sooner.

While sequencing costs continue to decline (currently costs are approximately \$300,000-\$500,000 to sequence the a human genome) and on the order of \$25 million for a 5X coverage of the genome, it is necessary that these cost the \$1,000 mark. Once this threshold has been reached it will be feasible for the majority of individuals to have their genomes encoded as part of their medical record.



Craig Venter

NEXT GENERATION SEQUENCING (NGS)

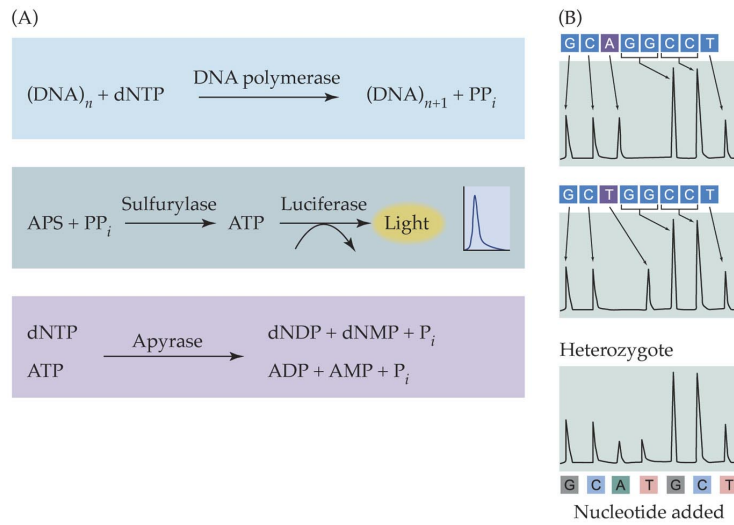
454 SEQUECING

ILLUMINA SEQUENCING

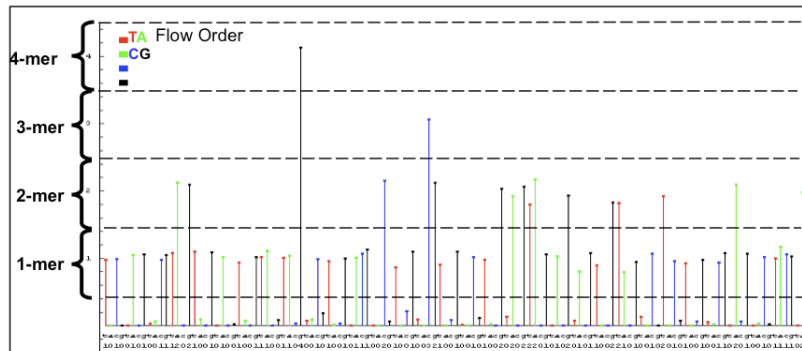
SOLID SEQUENCING

.....

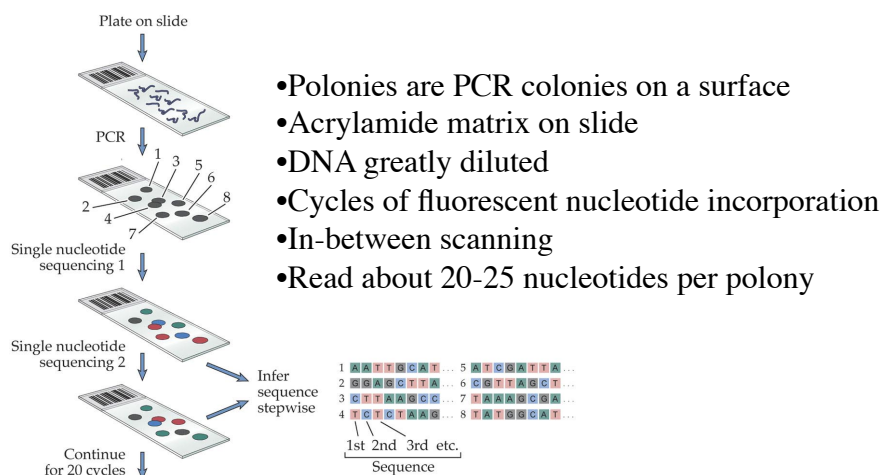
Important innovations I: Sequencing by light emission - Pyrosequencing



454 sequencer result



Important innovations II: Sequencing on surfaces - Polony sequencing



A PRIMER OF GENOME SCIENCE, Second Edition, Figure 2.8 © 2005 Sinauer Associates, Inc.

Illumina sequencer

illumina®

Log in to get personalized account information. Quick Order View Cart

858.202.4500 MyIllumina Tools

APPLICATIONS SYSTEMS CLINICAL SERVICES SCIENCE SUPPORT COMPANY

Search

Systems / HiSeq 2500/1500

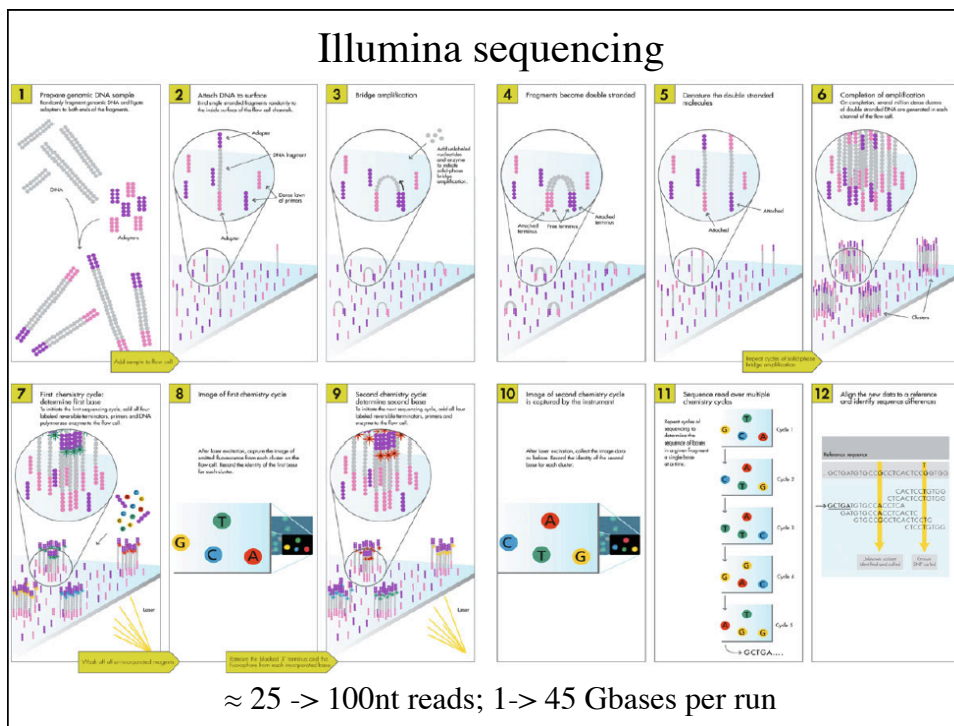
Subscribe | | Follow us:

Overview System Workflow Specs [GET A QUOTE](#)



Real flexibility.
Real throughput.
Real data quality.

The HiSeq 2500 is ready for any application,
any sample size—today.



Each Nation/University has its own sequencing centre

UNIVERSITY OF GÖTHEBURG
THE SAHLGRENKA ACADEMY

Genomics Core Facility

University of Gothenburg > The Sahlgrenska Academy > Core Facilities > Genomics Core Facility

Genomics Core Facility

Services provided

Access to Genomics

Instruments

Project request

Price list

Documents and protocols

Publications

Courses

Staff

Links

Quick links

- Metabarcoding/sequencing
- KAT500
- 7500 local computer
- NCBI tools
- DuPont
- MGC Genome Browser
- OSDP
- The Swedish project
- Applied Biosystems
- Illumina
- TruSight

News

New course at Genomics Core Facility

(15 Oct 2012) You can now apply for the course Data analysis in genomics given this spring by Genomics Core Facility.

Variant Minimization/RT-PCR course available on Geneanalyzer Forum

SciLifeLab

Genomics

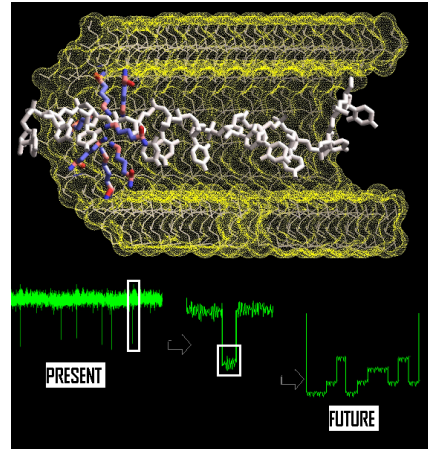
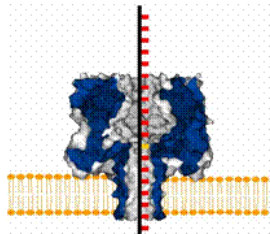
Genomics is the largest technical platform and provides access to technology for massively parallel-generation DNA sequencing and bioinformatics support. The technical development in the area has been overwhelming in recent years and the next generation sequencing instruments now allow large-scale genomics on a previously unattainable scale. Altogether 15 next generation sequencing instruments are available at present, with a combined capacity for DNA sequencing equal to several hundreds of complete human genomes per year. The next-generation DNA sequencing techniques can be used for a variety of studies, whole genome resequencing, complex genome sequencing (at specific sequences), de novo sequencing, targeted sequencing of regions in single or multiple individuals, transcriptome profiling including quantification and transcript isoforms and miRNAs, ChIP-Seq to detect transcription binding sites across the genome and targeted sequencing of amplicons such as 16 S rRNA genes and metagenomic sequencing of microbial genomes.

Modern genome analysis critically depend on expertise in computational biology (e.g. bioinformatics, statistics, and theoretical systems biology). Such expertise is closely integrated with the experimental genomics unit at SciLifeLab in order to optimize throughput, data handling, and analysis. As many researchers who will utilize the services at SciLifeLab do not have the experience required for efficient analysis of the data, this unit will provide a crucial bridge between the experimental platform and users.

SciLifeLab has been created by the coordinated effort of four universities in Stockholm and Uppsala: Stockholm University, the Karolinska Institute, The Royal Institute of Technology (KTH) and Uppsala University.

Partners: Karolinska Institutet, Stockholm University, Uppsala University.

For the "science fiction" future - Nanopore DNA sequencing

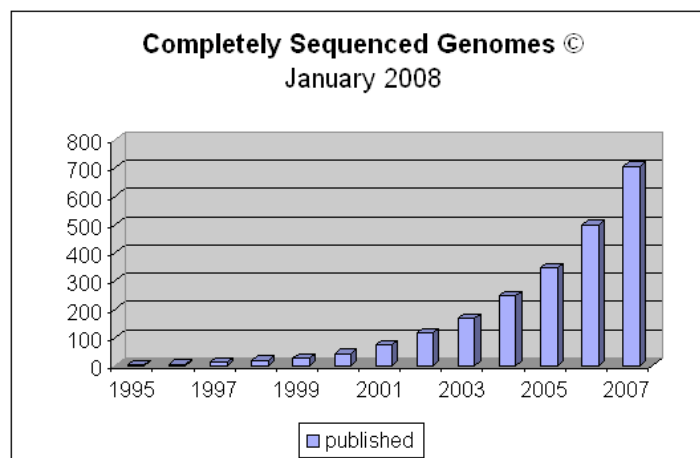


Summary sequencing

- Sanger sequencing has been extremely successful, but is currently a not so common way of DNA sequencing; still good since it provides long sequences.
- Several large centres in the world do the majority of sequencing projects – and is now being complemented by national/university centres.
- Who can reach the goal of the 1,000 dollar genome? We are in principle there.
- Soon reaching 1Gbase per hour.....or even faster....

FULLY SEQUENCED GENOMES - some landmark species

- 1995 - Haemophilus influenzae
- 1996 - Saccharomyces cerevisiae
- 1998 - Caenorhabditis elegans
- 2000 - Arabidopsis thaliana
- 2001 - Homo sapiens



March 2008 roughly 750 genomes
 March 2009 roughly 980 genomes
 March 2010 roughly 1220 genomes
 March 2011 roughly 1660 genomes
 March 2012 roughly 3000 genomes

Genomes On Line Database

currently (Feb 2013) 4132 completed genomes

GOLD Genomes Online Database

Genomes Online Database. Home

Last update: 2013-02-15
Total # of genomes: 21686

Home

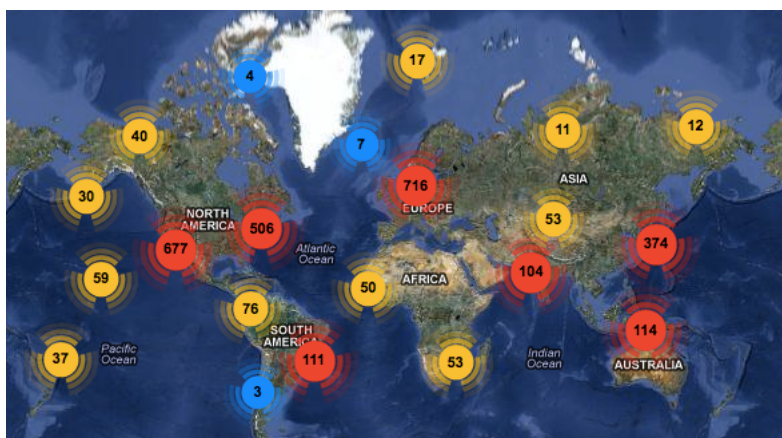
Genome Map
Genome Earth
Search
News
Statistics
Team

Welcome to the Genomes OnLine Database

GOLD-Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

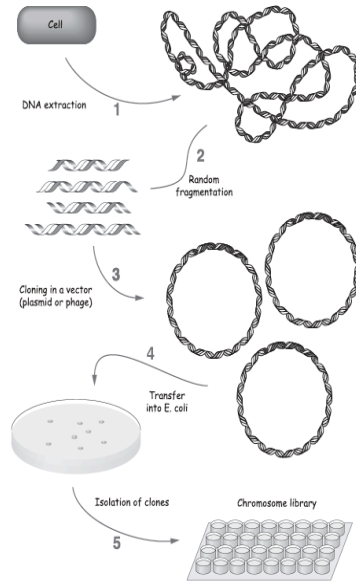
Metagenomes	Isolate Genomes	Genome Distribution
<ul style="list-style-type: none"> Classification Studies: 369 Samples: 2390 	<ul style="list-style-type: none"> Complete Projects: 4132 Incomplete Projects: 17514 Targeted Projects: 1077 	<ul style="list-style-type: none"> Project Type Sequencing Status Phylogenetic

Geography of finished genome projects

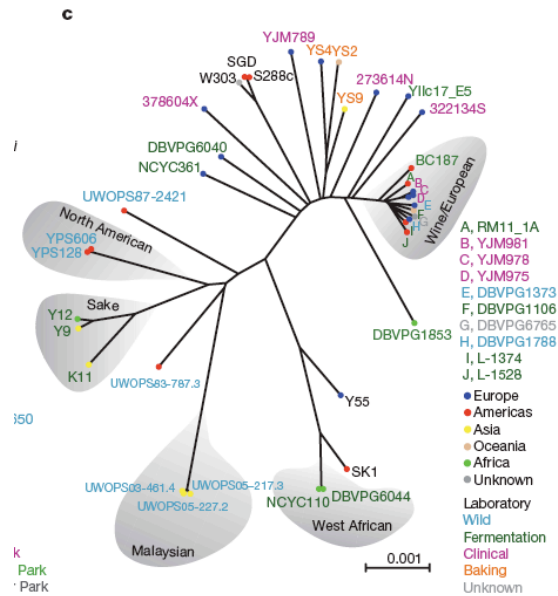


Genomes on line database Feb. 2013

Step I - Constructing a genome library



Who is being sequenced? Genome differences among *S. cerevisiae* strains

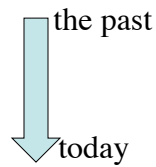


Genome difference between type strain S288c and the Malaysian strains is a roughly 1%; about the difference between humans and chimps

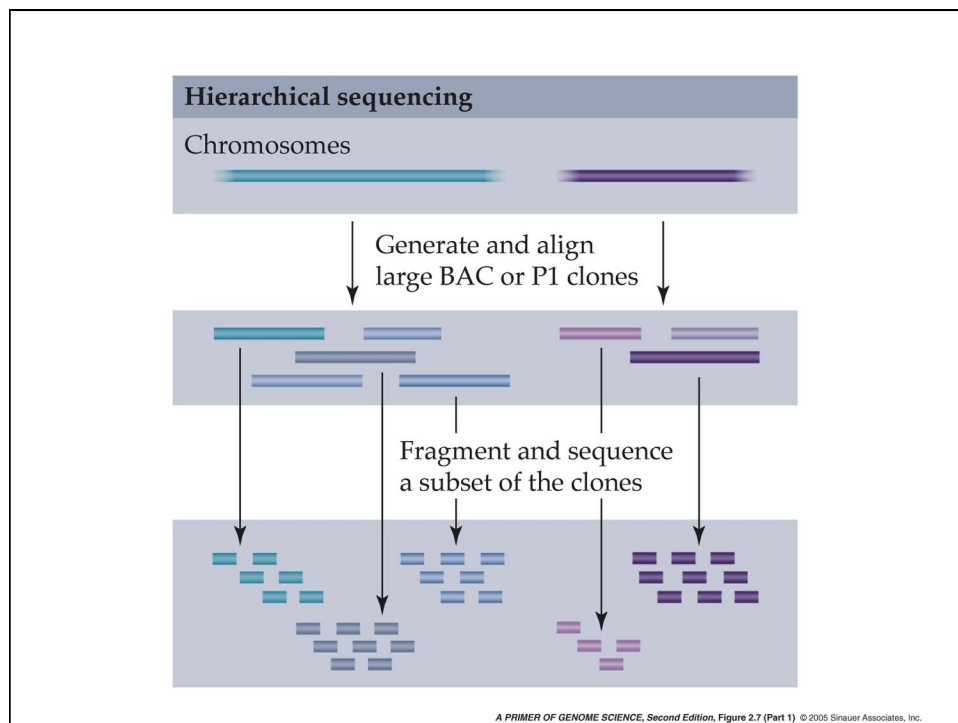
Liti et al., Nature 2009

VARIOUS SEQUENCING STRATEGIES

- Hierarchical sequencing: generate and align large BAC or YAC clones

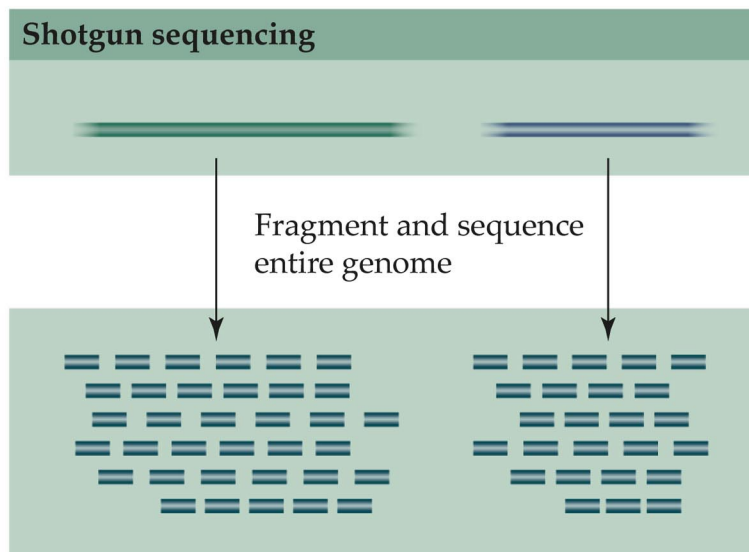


- Shotgun sequencing: fragment and sequence entire genome



II. SHOTGUN SEQUENCING

- Make fragment from whole genome
- Sequence a lot
- Align and make contigs in the computer



A PRIMER OF GENOME SCIENCE, Second Edition, Figure 2.7 (Part 2) © 2005 Sinauer Associates, Inc.

Forming contigs

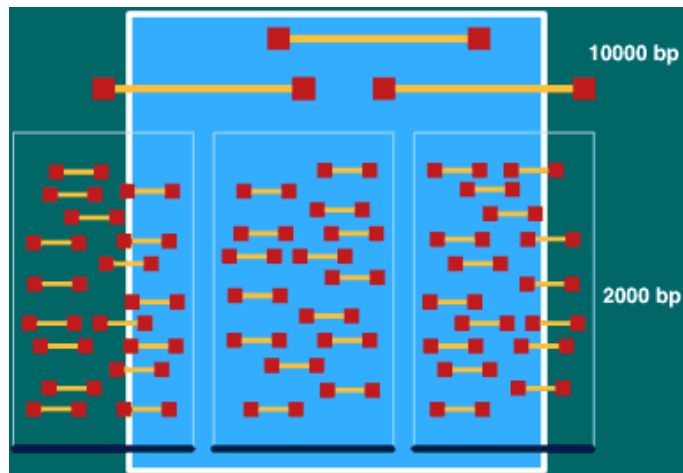


② homology with one contig - extension

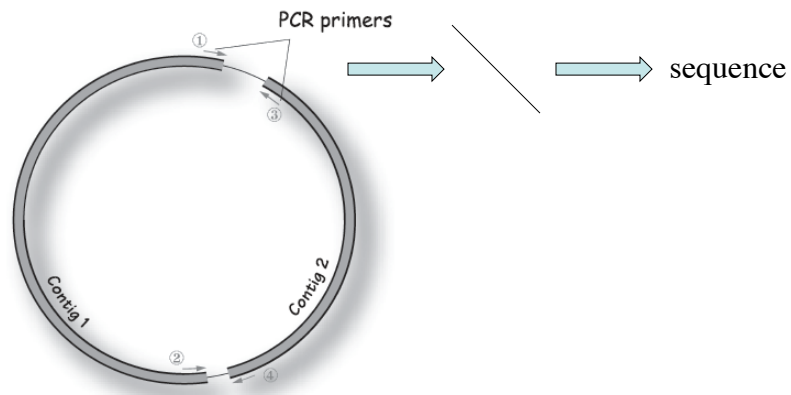


③ homology with two contigs - merge

Shotgun sequencing
- endsequencing of short and long inserts
form *contigs* and *scaffolds*



Filling gaps



When are we finished?

It is possible to estimate the amount of DNA that is sequenced as a function of fold coverage (Table 12.16). The probability a base is not sequenced was derived by Lander and Waterman (1988) and is given by

$$P_0 = e^{-c} \quad (12.1)$$

where c is the fold coverage and is given by

$$c = \frac{LN}{G} \quad (12.2)$$

and where LN is the number of bases sequenced, L being the read length and N the number of reads, and e is the constant 2.718. These results show that to achieve

TABLE 12-16 Probability That a Base Is Sequenced According To Equation 12.1

Fold Coverage	P_0	Percent Not Sequenced	Percent Sequenced
0.25	$e^{-0.25} = 0.78$	78	22
0.5	$e^{-0.5} = 0.61$	61	39
0.75	$e^{-0.75} = 0.47$	47	53
1	$e^{-1} = 0.37$	37	63
2	$e^{-2} = 0.135$	13.5	87.5
3	$e^{-3} = 0.05$	5	95
4	$e^{-4} = 0.018$	1.8	98.2
5	$e^{-5} = 0.0067$	0.6	99.4
6	$e^{-6} = 0.0025$	0.25	99.75
7	$e^{-7} = 0.0009$	0.09	99.91
8	$e^{-8} = 0.0003$	0.03	99.97
9	$e^{-9} = 0.0001$	0.01	99.99
10	$e^{-10} = 0.000045$	0.005	99.995

Source: Adapted from <http://www.genome.ou.edu/poisson.calc.html> and Lander and Waterman (1988).

Is *de novo* genome assembly using short reads possible?

- Jan 2010: Panda genome published (ca 2GB)
- 37 libs of sizes 150, 500, 2kb, 5kb, 10kb. [to handle repeats]
- Illumina, read length average 52 bp
- 176 GBases in total, 73x coverage
- Assembled with SOAPdenovo on 32 core 512GB computer

Vol 463 | 21 January 2010 | doi:10.1038/nature08696

nature

ARTICLES

The sequence and *de novo* assembly of the giant panda genome


Ruiqiang Li^{1,2*}, Wei Fan^{1*}, Geng Tian^{1,3*}, Hongmei Zhu^{1*}, Lin He^{4,5*}, Jing Cai^{3,6*}, Quanfei Huang¹, Qingle Cai^{1,7}, Bo Li¹, Yingqi Bai¹, Zhihe Zhang⁸, Yaping Zhang⁹, Wen Wang⁹, Jun Li¹, Fuwen Wei⁹, Heng Li¹⁰, Min Jian¹, Jianwen Li¹, Zhaolei Zhang¹¹, Rasmus Nielsen¹², Dawei Li¹, Wanjun Gu¹⁵, Zhentao Yang¹, Zhaoling Xuan¹, Oliver A. Ryder¹⁴, Frederick Chi-Ching Leung¹⁵, Yan Zhou¹, Jianjun Cao¹, Xiao Sun¹⁶, Yonggui Fu¹⁷, Xiaodong Fang¹, Xiaosen Guo¹

UNIVERSITY OF GOTHENBURG
CENTRE FOR MARINE EVOLUTIONARY BIOLOGY

Adjust | Listen | På svenska

Research Publications Activities People About CeMEB Outreach Contact us

University of Gothenburg > CeMEB



The Linnaeus Centre for Marine Evolutionary Biology

The Linnaeus Centre for Marine Evolutionary Biology, CeMEB, brings together a broad expertise in biology. We focus on evolutionary processes and mechanisms in marine species and populations. A main goal is to increase our understanding of how marine organisms adapt to new environmental conditions, for example changing sea water pH, temperature and salinity.

We started in July 2008, when we were selected for a ten year Linnaeus grant awarded by the Swedish Research Councils.

News & events


[New publication - Genetic Diversity and Ecosystem Functioning in the Face of Multiple Stressors](#)
[19 Sep 2012]

[New publication - Simulated climate change causes immune suppression and protein damage in the crustacean *Nephrops norvegicus*](#)
[19 Sep 2012]

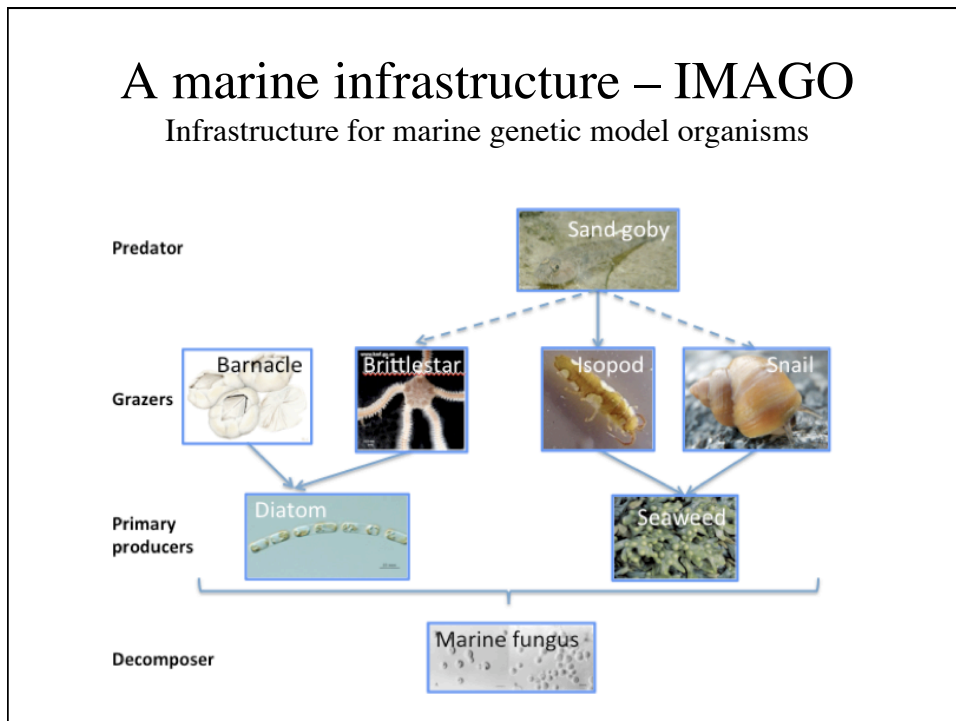
[New publication - Linking climate trends to population dynamics in the Baltic Ringed Seal: Impacts of historical and future winter temperatures](#)

Research for the future!

"Our research delivers urgently needed understanding of how marine organisms adapt to environmental changes over spatial and temporal scales relevant to current processes of global change."



Kerstin Johannesson, coordinator for the Linnaeus Centre for Marine Evolutionary Biology.





UNIVERSITY OF GOTHENBURG
CENTRE FOR MARINE EVOLUTIONARY BIOLOGY

IMAGO Marine Genome Project

Current status genome sequencing: 2012-10-10

CeMEB/IMAGO Organisms	Common name	Genome size (Gbp; haploid)	DNA libraries (fragment size)	Total number of Gbp	Total size assembly (Mbp)	contig N50	max contig size
<i>Balanus improvisus</i>	Bay barnacle	0.7 - 1.4	150, 300 & 3 000	180	509	1 514	61 346
<i>Amphiuira filiformis</i>	Brittlestar	2.5	300	34	811	822	216 866
<i>Debaryomyces hansenii</i> *	Marine yeast	0.0138	150	20	6-29	1 615 - 84 127	208 142 - 513 767
<i>Fucus vesiculosus</i>	Bladderwrack	1.1	300	34	176	453	64 287
<i>Idotea balthica</i>	Baltic isopod	2	300	45	771	695	402 660
<i>Littorina saxatilis</i>	Periwinkle	1.5	300 & 5 000	101	473	915	23 809
<i>Pomatoschistus minutus</i>	Sand goby	1	300	74	568	1 534	58 716
<i>Skeletonema marinoi</i>	Diatome	0.05-0.1		0	-	-	-

Total = 488Gbp



Diatome genomics

Surirella brebissonii

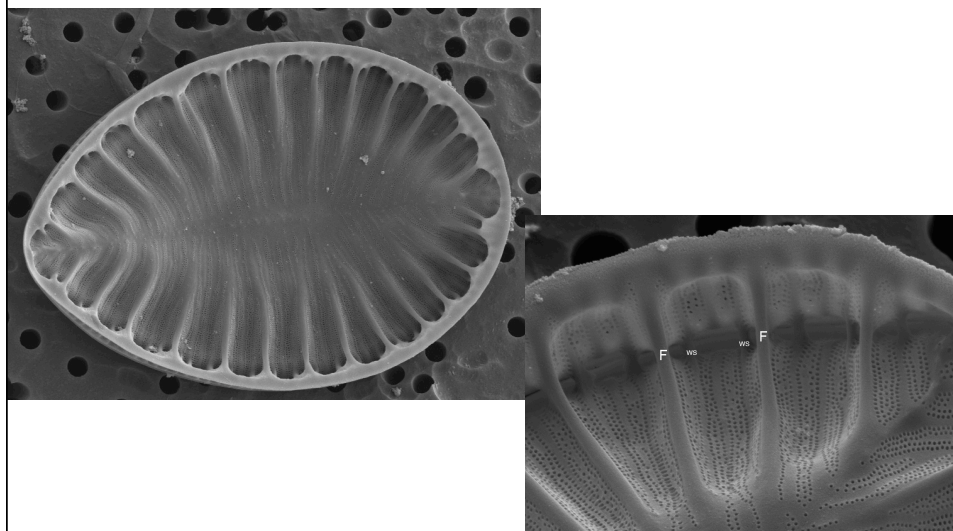


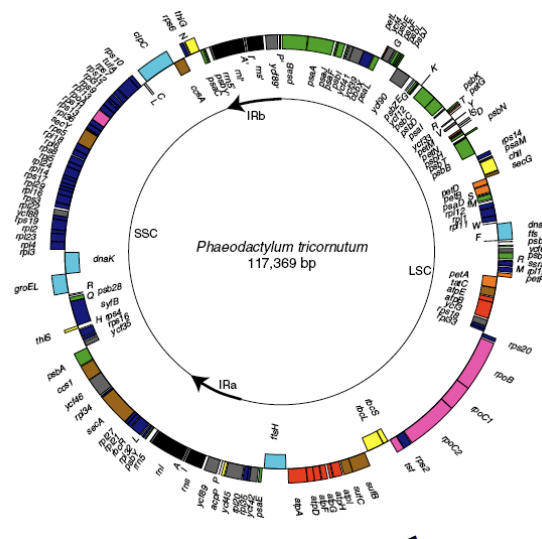
Photo Elisabeth Ruck

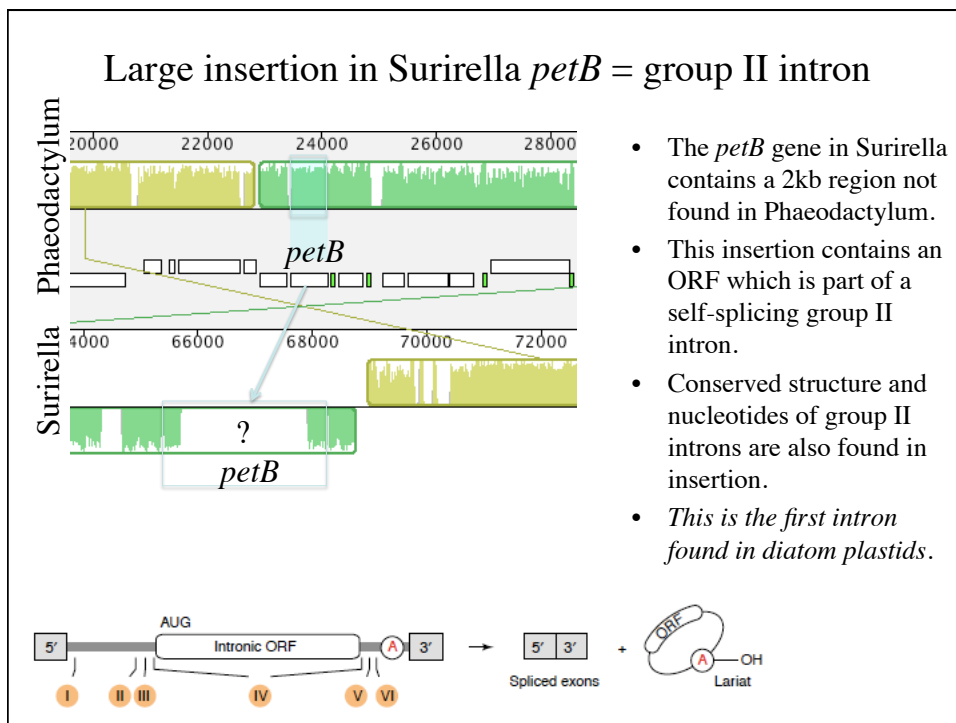
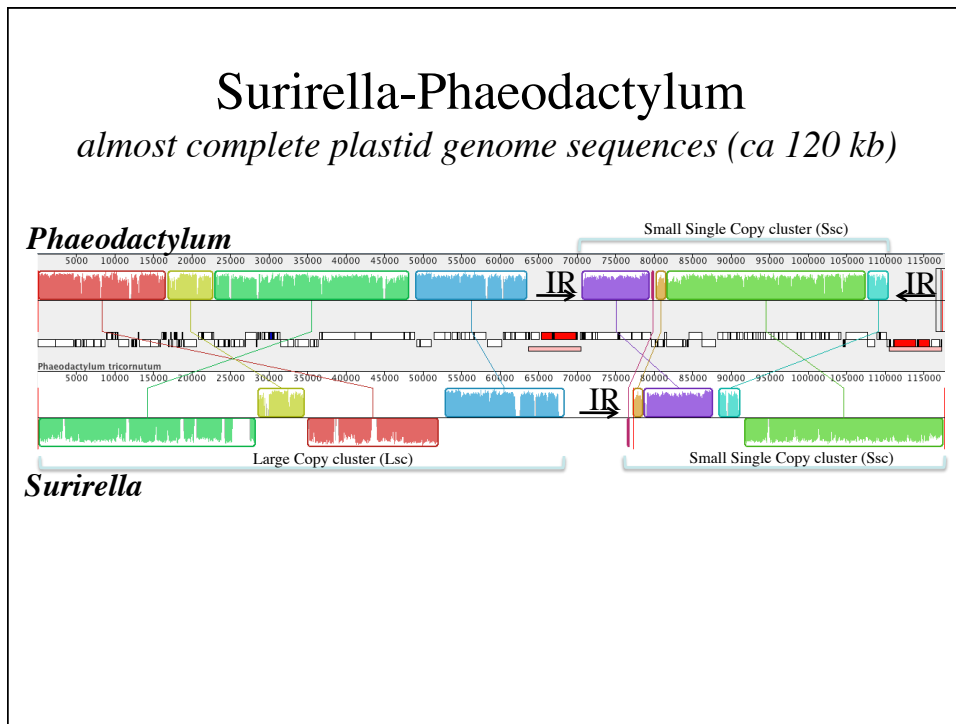
Surirella sequencing Illumina - HiSeq

Raw data from Surirella samples

- DNA 150 bp library, “paired-end” (2x100 nt, 50bp overlap)
125 million sequence pairs = **18 Gbases**
- DNA 3kb library, “mate-pair” (2x100 nt, 3kb gap)
106 million sequence pairs = **20 Gbases**
- RNA 300 bp library, “paired-end” (2x100 nt, 100 nt gap)
48 million sequence pairs = **10 Gbases** (24,000 contigs, >300 nt)

Chloroplast genome





Barnacle genomics

Balanus improvisus

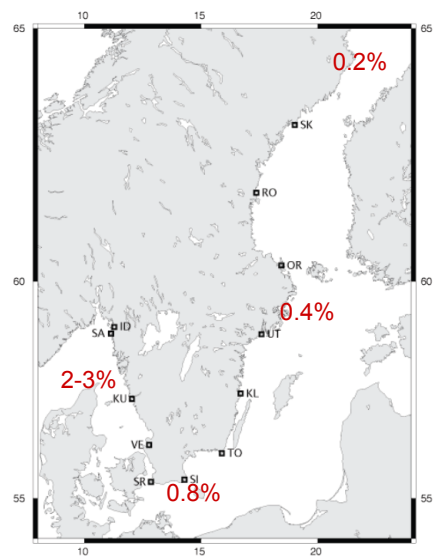


Why are barnacles interesting?

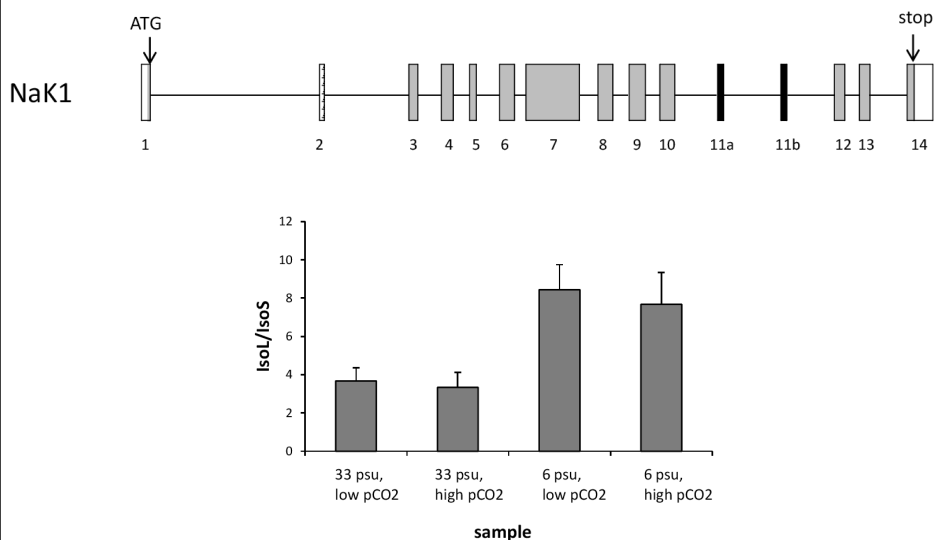
Biofouling organism



Adaptation to low salinities



Differential splicing of Na⁺/K⁺ ATPase (NaK)



GENOMICS AND SYSTEMS BIOLOGY

University of Gothenburg, Sweden

"Are you looking for an exciting and interdisciplinary education at the forefront of modern biology?
Here is the answer!"

MASTERS PROGRAMME

Practicals	Practicals	Practicals	Practicals
Project A	Project B	Project C	Project D
Advanced Functional Genomics (Anders Blomberg)	Advanced Bioinformatics (Magnus Alm Rosenblad)	Experimental Systems Biology (Stefan Hohmann)	Evolutionary Genomics (Jonas Warringer)
Proteomics/Metabolomics Phenomics Making models	Genome assembly RNA-seq Model extension	Single cell analysis Genetic interactions Dynamic modeling	Genome evolution Genotype to phenotype Modeling with SNPs

For more information and registration:

<http://www.science.gu.se/utbildning/masterprogram/program/Systembiologi>



UNIVERSITY OF GOTHENBURG