# Applied Metagenomics I

Till Helge Helwig

Eberhard-Karls-University Tübingen
Wilhelm-Schickard-Institute
Algorithms in Bioinformatics Group

Seminar “Metagenomics”

December 8th, 2009

# Introduction

## What we know already:

- **What** is metagenomics?
- **Sequencing** techniques
- Metagenome **analysis** with MEGAN

# Introduction

## What we know already:

- **What** is metagenomics?
- **Sequencing** techniques
- Metagenome **analysis** with MEGAN

## What I am going to explain:

- What is metagenomics **used** for?
- **Who** uses metagenomics?

# Overview

1. Applications overview
   - Bioprospecting
   - Phylogenetic Analysis
   - Functional Analysis
2. Complete Neanderthal Mitochondrial Genome
   - Introduction
   - Preparations & Procedures
3. Human Microbiome Project
   - Introduction
   - Methods & Project Status
4. Summary

# Overview

1. Applications overview
   - Bioprospecting
   - Phylogenetic Analysis
   - Functional Analysis
2. Complete Neanderthal Mitochondrial Genome
   - Introduction
   - Preparations & Procedures
3. Human Microbiome Project
   - Introduction
   - Methods & Project Status
4. Summary

# Basic Idea

- **New biomolecules** are required by different research fields, e.g.:
  - New agents are needed for **drug design**
  - Biocatalysts allow new **experimental protocols**

# Basic Idea

- **New biomolecules** are required by different research fields, e.g.:
  - New agents are needed for **drug design**
  - Biocatalysts allow new **experimental protocols**
- **Conservative search** is slow and has many manual steps, e.g.:
  - **Growing cultures** of selected microorganisms
  - **Selection** of new strains

# Basic Idea

- **New biomolecules** are required by different research fields, e.g.:
  - New agents are needed for **drug design**
  - Biocatalysts allow new **experimental protocols**
- **Conservative search** is slow and has many manual steps, e.g.:
  - **Growing cultures** of selected microorganisms
  - **Selection** of new strains

⇒ Instead of exploring single organisms let's look at **whole communities**

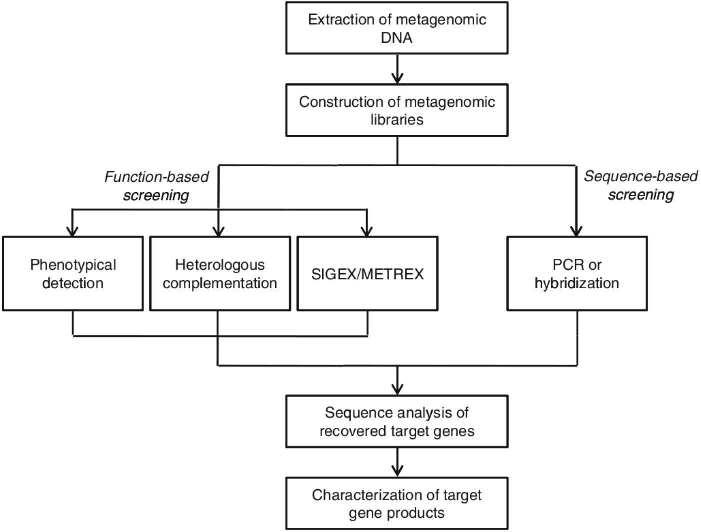
# Basic Idea

- **New biomolecules** are required by different research fields, e.g.:
  - New agents are needed for **drug design**
  - Biocatalysts allow new **experimental protocols**
- **Conservative search** is slow and has many manual steps, e.g.:
  - **Growing cultures** of selected microorganisms
  - **Selection** of new strains

⇒ Instead of exploring single organisms let's look at **whole communities**

⇒ **Increased chances** to be successful

# Overview



Extraction of metagenomic DNA

Construction of metagenomic libraries

*Function-based screening*

*Sequence-based screening*

Phenotypical detection

Heterologous complementation

SIGEX/METREX

PCR or hybridization

Sequence analysis of recovered target genes

Characterization of target gene products

Source: [Simon and Daniel, 2009]

# Sequence-Based Screening

- Uses **polymerase chain reaction** (PCR) or **hybridization**

# Sequence-Based Screening

- Uses **polymerase chain reaction** (PCR) or **hybridization**
- Requires **primers** obtained from known genes
- Identified genes have **similarity** with the reference genes
- Other genes are **not found**

# Sequence-Based Screening

- Uses **polymerase chain reaction** (PCR) or **hybridization**
- Requires **primers** obtained from known genes
- Identified genes have **similarity** with the reference genes
- Other genes are **not found**
- No dependency on **foreign host organisms**

# Sequence-Based Screening

- Uses **polymerase chain reaction** (PCR) or **hybridization**
- Requires **primers** obtained from known genes
- Identified genes have **similarity** with the reference genes
- Other genes are **not found**
- No dependency on **foreign host organisms**
- **Examples:**
  - “Subtractive hybridization magnetic bead capture”
  - “Metagenomic walking”
  - Microarrays

# Function-Based Screening

- Does **not** rely on available knowledge
- Can find **completely new** biomolecules
- Identifies only **complete genes** and not fragments

# Function-Based Screening

- Does **not** rely on available knowledge
- Can find **completely new** biomolecules
- Identifies only **complete genes** and not fragments
- Requires **foreign organisms** for **expression** of target genes and **production** of their proteins

# Function-Based Screening

- Does **not** rely on available knowledge
- Can find **completely new** biomolecules
- Identifies only **complete genes** and not fragments
- Requires **foreign organisms** for **expression** of target genes and **production** of their proteins
- **False-negative results** possible due to host's inability to adapt

# Function-Based Screening Methods

## Direct Detection

Phenotype identification by **indicators** within the growth medium.

# Function-Based Screening Methods

## Direct Detection

Phenotype identification by **indicators** within the growth medium.

## Heterologous Complementation

**Specific and highly selective medium** requires target genes to complement the organism's genes or host will not survive.

# Function-Based Screening Methods

## Direct Detection

Phenotype identification by **indicators** within the growth medium.

## Heterologous Complementation

**Specific and highly selective medium** requires target genes to complement the organism's genes or host will not survive.

## Induced Gene Expression

**Green fluorescent protein** is inserted together with the target gene via operon-trap expression vector. Relevant host cells are thus **visibly marked**.

# Screenings Summary

| | **Function-based** | **Sequence-based** |
|---|---|---|
| **Advantages** | • Only complete genes are found | • No need for a foreign host to obtain gene expression data |
| **Disadvantages** | • Relies on a foreign host, which might induce false negative results | • Cannot find entirely unknown genes<br>• Might yield incomplete genes |

# Who is out there?

- Explore the **phylogenetic diversity** within a sample
- Is also called “**taxonomical binning**”

# Who is out there?

- Explore the **phylogenetic diversity** within a sample
- Is also called "**taxonomical binning**"
- **Different approaches:**
  - Search for **known markers** (e.g. RecA)
  - Match reads against database and place them within a **taxonomy** (⇒ MEGAN)
  - Measure oligonucleotide or restriction-site **frequencies**
  - Compare and classify **16S rRNA** with the help of reference databases

# Who is out there?

- Explore the **phylogenetic diversity** within a sample
- Is also called "**taxonomical binning**"
- **Different approaches:**
  - Search for **known markers** (e.g. RecA)
  - Match reads against database and place them within a **taxonomy** (⇒ MEGAN)
  - Measure oligonucleotide or restriction-site **frequencies**
  - Compare and classify **16S rRNA** with the help of reference databases
- High potential for **inexact results** (e.g. due to PCR bias)

# Who is out there?

- Explore the **phylogenetic diversity** within a sample
- Is also called "**taxonomical binning**"
- **Different approaches:**
  - Search for **known markers** (e.g. RecA)
  - Match reads against database and place them within a **taxonomy** (⇒ MEGAN)
  - Measure oligonucleotide or restriction-site **frequencies**
  - Compare and classify **16S rRNA** with the help of reference databases
- High potential for **inexact results** (e.g. due to PCR bias)

⇒ **Shotgun sequencing** to avoid PCR

# What are they doing?

- Look at **functions** and **interactions** between microorganisms

# What are they doing?

- Look at **functions** and **interactions** between microorganisms
- **Functional databases** like SEED, Pfam and the STRING project provide **reference data**
- **Associate** sequences with these predefined clusters
- Also called "**functional binning**"

# What are they doing?

- Look at **functions** and **interactions** between microorganisms
- **Functional databases** like SEED, Pfam and the STRING project provide **reference data**
- **Associate** sequences with these predefined clusters
- Also called "**functional binning**"
- **Different organisms** can fulfill the same purpose
- The **same organism** can perform different tasks depending on the circumstances

# What are they doing?

- Look at **functions** and **interactions** between microorganisms
- **Functional databases** like SEED, Pfam and the STRING project provide **reference data**
- **Associate** sequences with these predefined clusters
- Also called "**functional binning**"
- **Different organisms** can fulfill the same purpose
- The **same organism** can perform different tasks depending on the circumstances
- **Tools** like MG-RAST are available already

# Overview

1. Applications overview
   - Bioprospecting
   - Phylogenetic Analysis
   - Functional Analysis
2. Complete Neanderthal Mitochondrial Genome
   - Introduction
   - Preparations & Procedures
3. Human Microbiome Project
   - Introduction
   - Methods & Project Status
4. Summary

# About the Project

- **Team** of 25 researchers...
- ...from institutes in the **USA** and **Europe**
- 38,000 years old **Neandertal bone** found in Vindjia Cave (Croatia)

Source: Wikipedia

# About the Project

- **Team** of 25 researchers...
- ...from institutes in the **USA** and **Europe**
- 38,000 years old **Neandertal bone** found in Vindjia Cave (Croatia)
- **Goal:** Finding new information about the **relationship** between modern humans and Neandertals

Source: Wikipedia

# Extraction and Preparation of the Sample

- Samples taken from a bone seem to be a **reliable source** for DNA

# Extraction and Preparation of the Sample

- Samples taken from a bone seem to be a **reliable source** for DNA
- **Contamination** with foreign DNA is possible due to previous washing procedures

# Extraction and Preparation of the Sample

- Samples taken from a bone seem to be a **reliable source** for DNA
- **Contamination** with foreign DNA is possible due to previous washing procedures
- **Specific primers** for human and Neandertal genes were searched
- PCR using these primers allowed for **quantification** of the contained DNAs

# Extraction and Preparation of the Sample

- Samples taken from a bone seem to be a **reliable source** for DNA
- **Contamination** with foreign DNA is possible due to previous washing procedures
- **Specific primers** for human and Neandertal genes were searched
- PCR using these primers allowed for **quantification** of the contained DNAs
- Contamination with unwanted modern human DNA was **below 1%**

# Considerations

- Ancient DNA is subject to **degradation processes**
- E.g. **deamination** of cytosine results in uracil residues, which are read as thymine by the DNA polymerase

# Considerations

- Ancient DNA is subject to **degradation processes**
- E.g. **deamination** of cytosine results in uracil residues, which are read as thymine by the DNA polymerase

## Considerations

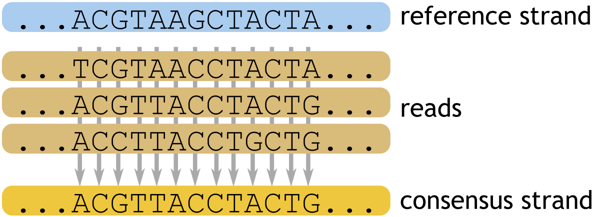
- Ancient DNA is subject to **degradation processes**
- E.g. **deamination** of cytosine results in uracil residues, which are read as thymine by the DNA polymerase
- Previous studies allowed **thorough understanding** of these disturbances

# Considerations

- Ancient DNA is subject to **degradation processes**
- E.g. **deamination** of cytosine results in uracil residues, which are read as thymine by the DNA polymerase
- Previous studies allowed **thorough understanding** of these disturbances
- To compensate for these expected problems **mitochondrial DNA** was chosen over nuclear DNA
- Each cell contains it in **huge abundance** and the **shorter length** works well with **454 sequencing**

# Assembly Process

- Nucleotide misincorporation is a **problem**
- Mitochondrial sequence from modern humans used as **reference strand**



...ACGTAAGCTACTA... reference strand

...TCGTAACCTACTA...
...ACGTTACCTACTG...
...ACCTTACCTGCTG...
reads

...ACGTTACCTACTG... consensus strand

- Sequencing reads **aligned** with the reference
- **Majority base** identified for each column

# Assembly Process (2)

- Some regions were **problematic** due to e.g. missing coverage
- These were **extracted specifically** from another bone sample and Sanger **sequenced**

# Assembly Process (2)

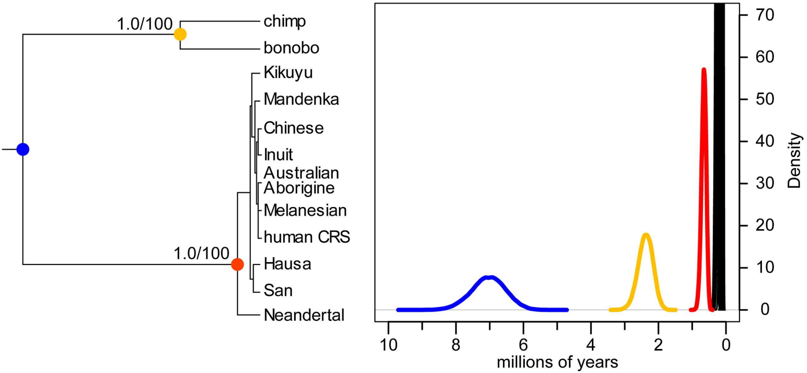
- Some regions were **problematic** due to e.g. missing coverage
- These were **extracted specifically** from another bone sample and Sanger **sequenced**
- After **repairing** the consensus strand using those results the **new** consensus strand was used as reference strand
- 721 sequences **additional** sequences were found, which the first step did not reveal

# Results

- A total of **8341 sequences** could be identified
- This leads to a **34.9-fold coverage** of the whole mitochondrial genome
- **Verification** steps showed a contamination with modern human mtDNA of 0.5%
- Trusting this to be fairly reliable the mtDNA was **analyzed** and **compared** with other data

# Results (2)

- Thus a **phylogenetic tree** could be estimated



Source: [Green et al., 2008]

# Results (3)

- The Neandertal mitochondrial genome is definitely **no mere variation** of the modern human's version
- About **660,000 years ago** both lineages diverged
- Their most recent common ancestor lived quite some time **before** the most recent common ancestor of all humans
- The results also suggest that the Neandertal **population size** was significantly **smaller** than the modern ones

# Overview

1. Applications overview
   - Bioprospecting
   - Phylogenetic Analysis
   - Functional Analysis
2. Complete Neanderthal Mitochondrial Genome
   - Introduction
   - Preparations & Procedures
3. Human Microbiome Project
   - Introduction
   - Methods & Project Status
4. Summary

# About the Project

- **Whole human genome** published in 2003

# About the Project

- **Whole human genome** published in 2003
- This is **not** the only **genetic information** associated with humans

# About the Project

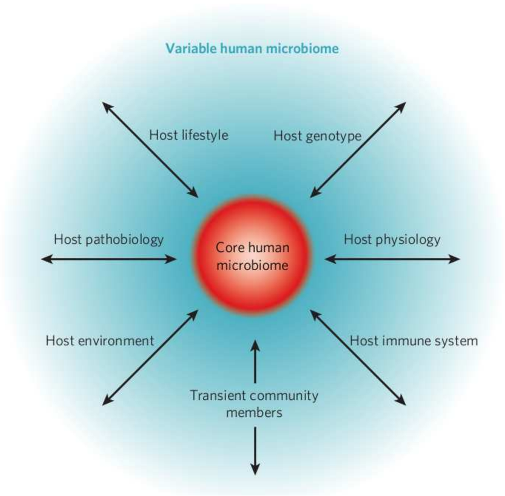
- **Whole human genome** published in 2003
- This is **not** the only **genetic information** associated with humans
- Constant **symbiosis** with a vast number of **microorganisms** (microbiota)
- They perform tasks we therefore **never** had to do ourselves

# About the Project

- **Whole human genome** published in 2003
- This is **not** the only **genetic information** associated with humans
- Constant **symbiosis** with a vast number of **microorganisms** (microbiota)
- They perform tasks we therefore **never** had to do ourselves
- **Goal:** Characterize the distribution and evolution of microbiota

# Microbiome

- Entirety of **all** microbiota genomes
- HMP defines **core** and **variable microbiome**



Source: [Turnbaugh et al., 2007]

# Questions

- Is there a **core microbiome**?
- Do all humans have the **same** core microbiome?

# Questions

- Is there a **core microbiome**?
- Do all humans have the **same** core microbiome?
- Which **factors** influence the variable microbiome?
- How **stable** is the microbiome?

# Questions

- Is there a **core microbiome**?
- Do all humans have the **same** core microbiome?
- Which **factors** influence the variable microbiome?
- How **stable** is the microbiome?
- Is **manipulation** of the microorganisms possible to increase their **performance**?

# Questions

- Is there a **core microbiome**?
- Do all humans have the **same** core microbiome?
- Which **factors** influence the variable microbiome?
- How **stable** is the microbiome?
- Is **manipulation** of the microorganisms possible to increase their **performance**?
- How do the microbiota relate to certain **diseases**?

# Reference Database

- **Metagenomic methods** will be applied to samples taken from human individuals
- These rely on **reference data**

# Reference Database

- **Metagenomic methods** will be applied to samples taken from human individuals
- These rely on **reference data**
- Thus the first step is the creation of a suitable **database** containing at least **1000 relevant genomes**
- They are chosen by information from **16S-rRNA-gene-based surveys**

# Reference Database

- **Metagenomic methods** will be applied to samples taken from human individuals
- These rely on **reference data**
- Thus the first step is the creation of a suitable **database** containing at least **1000 relevant genomes**
- They are chosen by information from **16S-rRNA-gene-based surveys**
- For each of the selected organisms DNA has to be **acquired**
- Many of the microorganisms can **not** be cultured

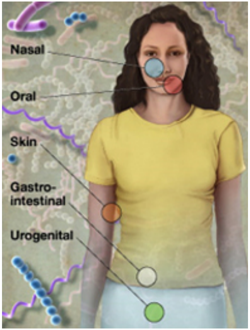
# Reference Database

- **Metagenomic methods** will be applied to samples taken from human individuals
- These rely on **reference data**
- Thus the first step is the creation of a suitable **database** containing at least **1000 relevant genomes**
- They are chosen by information from **16S-rRNA-gene-based surveys**
- For each of the selected organisms DNA has to be **acquired**
- Many of the microorganisms can **not** be cultured

⇒ Immense **community effort**

# Fields of Interest

- **Five representative habitats** were chosen for analysis
  - Nasal
  - Oral
  - Skin
  - Gastrointestinal
  - Urogenital
- Samples from each will be analyzed once the reference data is **complete**



Source: http://www.hmpdacc-resources.org

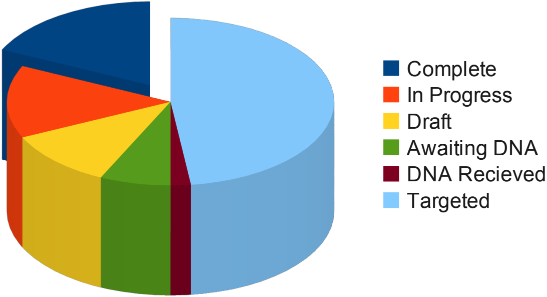
# Metagenomics Relevance

- The project will generate **huge amounts of data**
- Fast and easy methods to **manage** and **access** them have been and will be explored

# Metagenomics Relevance

- The project will generate **huge amounts of data**
- Fast and easy methods to **manage** and **access** them have been and will be explored
- Reads from whole-genome shotgun sequencing will be **sorted** by species or at least taxonomical groups
- **Building** and **handling** phylogenetic trees containing **millions** of sequences will have be optimized

# Current Status



Source: http://www.hmpdacc-resources.org

- 18% of the reference genomes **completed**
- The remaining ones in different states of **preparation** or **precessing**

# Overview

1. Applications overview
    - Bioprospecting
    - Phylogenetic Analysis
    - Functional Analysis
2. Complete Neanderthal Mitochondrial Genome
    - Introduction
    - Preparations & Procedures
3. Human Microbiome Project
    - Introduction
    - Methods & Project Status
4. Summary

# What You Should Take Home

- The number of **possible applications** for metagenomics is immense
- The spectrum reaches from **narrowing down** on one **specific** genome to looking at a **vast number** of organisms **at once**
- Due to fast growing projects with **increasing needs** for efficient methods the field of metagenomics will keep **growing** fast
- You definitely have not heard the last of **applied metagenomics** and **metagenomics in general**

**Thank you very much for your attention.**

Questions? Remarks?

# References I

Blow, N. (2008).
Metagenomics: exploring unseen communities.
*Nature*, 453(7195):687.

Green, R., Malaspinas, A., Krause, J., Briggs, A., Johnson, P., Uhler, C., Meyer, M., Good, J., Maricic, T., Stenzel, U., et al. (2008).
A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing.
*Cell*, 134(3):416–426.

Huson, D., Auch, A., Qi, J., and Schuster, S. (2007).
MEGAN analysis of metagenomic data.
*Genome research*, 17(3):377.

LeCleir, G., Buchan, A., and Hollibaugh, J. (2004).
Chitinase gene sequences retrieved from diverse aquatic habitats reveal environment-specific distributions.
*Applied and environmental microbiology*, 70(12):6977.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008).
The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes.
*BMC bioinformatics*, 9(1):386.

Meyer, Q., Burton, S., and Cowan, D. (2007).
Subtractive hybridization magnetic bead capture: A new technique for the recovery of full-length ORFs from the metagenome.
*Biotechnology Journal*, 2(1):36.

# References II

Morimoto, S. and Fujii, T. (2009).
A new approach to retrieve full lengths of functional genes from soil by PCR-DGGE and metagenome walking.
*Applied Microbiology and Biotechnology*, 83(2):389–396.

Pathak, G., Ehrenreich, A., Losi, A., Streit, W., and G
"artner, W. (2009).
Novel blue light-sensitive proteins from a metagenomic approach.
*Environmental Microbiology*, 11(9):2388–2399.

Richter, D., Ott, F., Auch, A., Schmid, R., and Huson, D. (2008).
Metasim—a sequencing simulator for genomics and metagenomics.
*PLoS One*, 3(10).

Simon, C. and Daniel, R. (2009).
Achievements and new knowledge unraveled by metagenomic approaches.
*Applied Microbiology and Biotechnology*, pages 1–12.

Turnbaugh, P., Ley, R., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. (2007).
The human microbiome project.
*Nature*, 449(7164):804–810.

Von Mering, C., Hugenholtz, P., Raes, J., Tringe, S., Doerks, T., Jensen, L., Ward, N., and Bork, P. (2007).
Quantitative phylogenetic assessment of microbial communities in diverse environments.
*Science*, 315(5815):1126.