Personalized Health

cb2_intro_personalhealth

Protein Prediction 2 - Protein function
Computational Biology 1
TUM Winter 2014/15
Videos: YouTube / www.rostlab.org

THANKS:
Tim Karl + Jonas Reeb

Special lectures:
• xxx - Tobias Hamp

No lecture:
• Nov 12 Tue (Student assembly)
• Dec 12 Thu (TUM Dies Academicus)

LAST lecture: January 20

Examen: January 22
• Makeup: Apr 14, 2015 - morning/noon
# Lecture plan (CB2 function)

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>01: 2014/10/07</td>
<td>no lecture</td>
</tr>
<tr>
<td>02: 2014/10/09</td>
<td>welcome: who we are</td>
</tr>
<tr>
<td>03: 2014/10/14</td>
<td>no lecture (prof sick)</td>
</tr>
<tr>
<td>04: 2014/10/16</td>
<td>no lecture (prof sick)</td>
</tr>
<tr>
<td>05: 2014/10/21</td>
<td>Personalized medicine - predict effects of SNPs</td>
</tr>
<tr>
<td>06: 2014/10/23</td>
<td>Intro - function 1: concepts / homology</td>
</tr>
<tr>
<td>07: 2014/10/28</td>
<td>Tobias Hamp: Homology-based prediction of function</td>
</tr>
<tr>
<td>08: 2014/10/30</td>
<td>Tobias Hamp: Homology-based prediction of function 2</td>
</tr>
<tr>
<td>09: 2014/11/04</td>
<td>no lecture: SVV (student reps)</td>
</tr>
<tr>
<td>10: 2014/11/06</td>
<td>Intro - function 3: motifs</td>
</tr>
<tr>
<td>11: 2014/11/11</td>
<td>Localization 1</td>
</tr>
<tr>
<td>12: 2014/11/13</td>
<td>Localization 2</td>
</tr>
<tr>
<td>13: 2014/11/18</td>
<td>Localization 3 - Tatyana Goldberg</td>
</tr>
<tr>
<td>14: 2014/11/20</td>
<td>Protein-protein interaction 1</td>
</tr>
<tr>
<td>15: 2014/11/25</td>
<td>Protein-protein interaction 2</td>
</tr>
<tr>
<td>16: 2014/11/27</td>
<td>Protein-DNA/RNA interaction</td>
</tr>
<tr>
<td>17: 2014/12/02</td>
<td>SNP effect 1</td>
</tr>
<tr>
<td>18: 2014/12/04</td>
<td>no lecture: Dies Academicus</td>
</tr>
<tr>
<td>19: 2014/12/09</td>
<td>SNP effect 2</td>
</tr>
<tr>
<td>20: 2014/12/11</td>
<td>SNP effect 3 / Marco De Vivo (ISS Genoa) - Drug Design</td>
</tr>
<tr>
<td>21: 2014/12/16</td>
<td>Andrea Schafferhans: 3D function prediction</td>
</tr>
<tr>
<td>22: 2014/12/18</td>
<td>Andrea Schafferhans: Docking</td>
</tr>
<tr>
<td>23-26:</td>
<td>no lectures - winter break (2014/12/24 - 2015/01/06)</td>
</tr>
<tr>
<td>27: 2015/01/08</td>
<td>Punta - Pfam</td>
</tr>
<tr>
<td>28: 2015/01/13</td>
<td>Marco De Vivo (ISS Genoa) - Drug Design</td>
</tr>
<tr>
<td>29: 2015/01/15</td>
<td>GO enrichment</td>
</tr>
<tr>
<td>30: 2015/01/20</td>
<td>WRAP up !Protein-DNA/RNA interaction 2</td>
</tr>
<tr>
<td>31: 2015/01/22</td>
<td>examen</td>
</tr>
</tbody>
</table>
Today: Intro into function

LAST WEEKs
- Intro into lecture

THIS WEEK
- Tuesday: From molecular predictions to individualized medicine
- Introduction to protein function I

NEXT WEEK
- Tobias Hamp: function prediction
Beaming from observed SNPs to individual health

“Beam me up, Scotty!”

William Shatner as James Kirk
Leonard Nimoy as Spock
(Starship Enterprise)

© Wikipedia
I1: Individualized medicine / personalized medicine
Map genotype to phenotype
Personalized medicine - SNPs


Slide: © Marc Offman (TUM Munich)
Connecting change and computer


Slide: © Marc Offman (TUM Munich)
PredictProtein

Laszlo Kajan

Guy Yachd
From molecular to system


Slide: © Marc Offman (TUM Munich)
I2: human genome completely sequenced
Sequencing complete organisms
Central dogma of molecular biology

DNA
information, code, library, manual

RNA
intermediate step

Protein
machinery of life

© Burkhard Rost
Haemophilus Influenzae (Pfeiffer’s bacillus) 40 authors

Mycoplasma Genitalium

Saccharomyces cerevisiae

Caenorhabditis elegans, nematode - worm

Drosophila melanogaster  fruit fly 195 authors
## Genome sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>470</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>1740</td>
</tr>
<tr>
<td>Methanococcus jannaschi</td>
<td>1,738.00</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4,288.00</td>
</tr>
<tr>
<td>Sacharomyces cerevisiae - yeast</td>
<td>6,600.00</td>
</tr>
<tr>
<td>Drosophila melanogaster - fruit fly</td>
<td>13,600.00</td>
</tr>
<tr>
<td>Caenorhabditis elegans - worm</td>
<td>19,000.00</td>
</tr>
<tr>
<td>Arabidopsis thaliana - mustard</td>
<td>26,735.00</td>
</tr>
<tr>
<td>Oryza sativa - rice</td>
<td>50,000.00</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>*100,000</td>
</tr>
</tbody>
</table>

* Estimate from 1999
not Jan 1, 2000

number of genes/proteins:
Oct 1999
(after >5 years):
100,000

Nov 1999:
oops there are only 30,000
# Genome sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>470</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>1740</td>
</tr>
<tr>
<td>Methanococcus jannaschi</td>
<td>1,738.00</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4,288.00</td>
</tr>
<tr>
<td>Sacharomyces cerevisiae - yeast</td>
<td>6,600.00</td>
</tr>
<tr>
<td>Drosophila melanogaster - fruit fly</td>
<td>13,600.00</td>
</tr>
<tr>
<td>Caenorhabditis elegans - worm</td>
<td>*21,000</td>
</tr>
<tr>
<td>Arabidopsis thaliana - mustard</td>
<td>*26,735</td>
</tr>
<tr>
<td>Oryza sativa - rice</td>
<td>*50,000</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>*25,000</td>
</tr>
</tbody>
</table>

* Estimates from 2010
early idea: 
we need one human genome
23 chromosome pairs: 3.4G base pairs

Spectral karyogram of human female

© Wikipedia
Genome sizes (in basepairs)

- Mycoplasma
- Gram positive bacteria
- Gram negative bacteria
- Fungi / Moulds
- Algae
- Worms
- Crustaceans
- Echinoderms
- Insects
- Mollusks
- Birds
- Bony fish
- Cartilaginous fish
- Reptiles
- Mammals
- Amphibians
- Flowering Plants

(in bp)

© Burkhard Rost
© Wikipedia
23 chromosome pairs: 3.4G base pairs

coding: <1.5%

how many genes/proteins?
Central dogma of molecular biology

DNA → RNA → Protein

- DNA: information, code, library, manual
- RNA: intermediate step
- Protein: machinery of life

© Burkhard Rost
FANTOM 3: Functional annotation of mouse

The discovery of a novel RNA Continent through comprehensive analyses of mammalian transcriptomes

FANTOM Consortium

-63

e of the Mammalian Genome

143 authors
Map still not fully complete?

- 2011:
  - 20-25,000 proteins
- considerable alternative splicing
- considerable amount of non-coding RNA
now that we have the parts, do we understand life?
“Like for every good manual: you hardly ever find what you look for when you find it, it is difficult to understand”
Anna Tramontano, La Sapienza, Rome, Italy

Now we know it all?

http://static.open.salon.com
http://i42.tinypic.com
Can we compute life?

- Tough challenge, we address it from all possible angles
- But
  - We still do not understand the “book”/genome
  - We don’t even understand the first principles, completely
- Nevertheless
New life in the test tube!

first synthesis of an entire organism

Mycoplasma genitalium

Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome.

We can dream but we cannot compute life, yet
Mapping genome diversity
Genes to proteins

<table>
<thead>
<tr>
<th>DNA / Genes</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>information, library.</td>
<td>machinery of life</td>
</tr>
<tr>
<td>human genome: 3 billion letters</td>
<td>human proteome:</td>
</tr>
</tbody>
</table>

| 4000 books |

human genome: 3 billion letters
human proteome: ~20 thousand proteins

![DNA structure](image1.png)
![Structure of proteins](image2.png)
You hardly ever find what you look for.

When you find it, it is difficult to understand.
HapMap project (haplotype map):
- common genetic variation in homo sapiens

1000 Genomes Project
- launched 2008
- NIH - NCHGR: National Center for Human Genome Research
- Sanger/EBI
- BGI/Shenzen

10k UK
1m chinese

© Wikipedia
How much do we differ?
we differ by 20,000 amino acids (letters)

- same letter
- different letter
most cells in you have the same library!

4000 books

I may have 55 trillion cells *
today: age of diversity and variation
Epigenetics

- heritable changes not related to DNA
- Greek: epi (ἐπί): (above|over|outer)

<table>
<thead>
<tr>
<th>EPIGENETIC MECHANISMS</th>
<th>HEALTH ENDPOINTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development (in utero, childhood)</td>
<td>Cancer</td>
</tr>
<tr>
<td>Environmental chemicals</td>
<td>Autoimmune disease</td>
</tr>
<tr>
<td>Drugs/Pharmaceuticals</td>
<td>Mental disorders</td>
</tr>
<tr>
<td>Aging</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Diet</td>
<td></td>
</tr>
</tbody>
</table>

DNA + protein (histones) → chromatin

© Wikipedia
Discover diversity and movement

“migration from Siberia into New World ... 5,500 years ago, independent of ... Native Americans and Inuit”
Who Were the Denisovans?

Science 26 August 2011: vol. 333 no. 6046 1084–1087

Geneic history of an archaic hominin group from Denisova Cave in Siberia

David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin & Svante Pääbo

Affiliations | Contributions | Corresponding authors

Nature 468, 1053–1060 (23 December 2010) doi:10.1038/nature09710
Received 15 August 2010 | Accepted 30 November 2010 | Published online 22 December 2010

Abstract

Abstract • Introduction • DNA sequence determination • Human DNA contamination estimates • Ancestral features and duplications • Relationship to Neanderthals and modern humans • A Neanderthal-specific bottleneck • No Denisovan gene flow into all Eurasians • Denisovan gene flow into the ancestors of Melanesians • A model of population history • Discordance of mtDNA and nuclear histories • A tooth from Denisova Cave • Morphology of the Denisova molar • Stratigraphy and dating • Discussion • Methods • Accession codes • References • Acknowledgements • Author information • Supplementary information • Comments
Human Diversity


http://image1.masterfile.com/em_w/01/30/11/619-01301192w.jpg
Do people differ in their genes?

they differ little?  they differ more?  they are similar?

Sweden  Italy  China

South Africa  Bushmen

http://www.china.com.cn/attachement/jpg/site1007/20080808/00105cad26fb0a0635a710.jpg


© Burkhard Rost

MACMILLAN PUBLISHERS LIMITED. ALL RIGHTS RESERVED
Discover diversity and movement

“Bushmen ... more different from each other than ... European and Asian”

“migration from Siberia into New World ... 5,500 years ago, independent of ... Native Americans and Inuit”

© Burkhard Rost
ROSTLAB.
I3: From gene to protein variation
SNPs are changes of one single nucleotide / letter

SNPs are changes of one single nucleotide / letter

© Wikipedia
change of amino acid: non-synonymous (ns)
ocurrence > T% in population: SNP

here:

ANY amino acid change
-> nsSNP (or SNP)
nsSNPs importantly involved in disease
Individual medicine: make it relevant

Validated human SNPs in dbSNP

Figs. 2 - 1
remember: more than genome:
Epi-genetics & copy number variation & many more
Get disease high-throughput manner?

☐ GWAS (Genome Wide Association Study)
☐ many times (patients, drugs)
☐ environmental factors (virus/bacteria)
☐ map all this to interactions (of genes/proteins/drugs/environment)
☐ put into simple pipeline that is readily available for clinicians
☐ job done?
Many challenges
top-down: killing power of numbers
(many people -> loads of data)
Illumina: HiSeq (2000)

- run: 1-8 days, 24 GB/day
- 1 human genome @ 30x / day
- full capacity: 25 TB data / day
- 120 x 76 x 94 cm

BGI - Shenzhen
Ion Torrent: Benchtop Ion Proton

- $1000 human genome
- run: 6 hours
- full capacity: ~5-10

Ion Torrent - Jan 2012
Illumina: MiSeq

- run: 6 hours
- full capacity: ~5-10 TB data / day

Illumina - Early 2012
Michelangelo Buonarroti (1475-1564)

www.webexhibits.org/colorart/i/michelangelo-creation-adam-.jpg
Central dogma of molecular biology

DNA: information, code, library, manual
RNA: intermediate step
Protein: machinery of life

Central Dogma: Francis Crick, Marshall Nirenberg
top-down: communication

Understand peers
bottom-up: annotation
function known for 10–50% of human
function known for 10-50% of human
function known for 10-50% of human
function known for 10-50% of human

http://i42.tinypic.com
function known for 10-50%
Rostlab & friends @ ISMB/ECCB Vienna
Our niche: evolution + machine learning

Sequence $\rightarrow$ PSI-BLAST $\rightarrow$ Filter

MaxHom

PHDsec

1993

60% $\rightarrow$ 72% / 77%

B Rost 1996 Meth Enzymol 266:525-539
Exciting projects

- LOCtree etc: predict localization
- Predict enzymatic activity & flexibility
- Protein disorder
- Predict membrane regions, epitomes,
- Improve alignment methods
- SNP-pipeline: predict nsSNP effects
- PredictProtein: web service since 1992
- NESG & NYCOMPS: structural genomics
Personalized medicine vs personalized health
person-specific drug development realistic dream?
Drug development: time

Total time (on average) = ~13.5 years

© Marco De Vivo, IIT Genova

Drug development: time

~$820m to discover (~6 years)

~$960m to develop (~7.5 years)

Total costs (on average) = ~$1.8billion

~13.5 years

© Marco De Vivo, IIT Genova

< 40 NMEs approved annually by FDA

NME = new molecular entity (new drug)

Using: M Allison (2012) Nat Biotech 30, 41-49 doi:10.1038/nbt.2083 Fig. 2
personalized health & diagnosis & administration

not

personalized drugs
Predict nsSNP effects
SNPs are changes of one single nucleotide / letter

SNPs are changes of one single nucleotide / letter

© Wikipedia
Big changes may not matter!

Sequence identity implies structural similarity!

Don't know region

Percentage sequence identity

Number of residues aligned

Distance from curve = +10

Distance from curve = -10

C Sander & R Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94
Yana Bromberg, Rutgers University
nsSNP effects: some in silico methods

SIFT
PC Ng & S Henikoff (2003) NAR 31:3812-14
>Sequence
VHLTP EKSA VTALWGKVN V FFESFGDLST PDAVMGNPKV
KAHGKKVLGA
Mutant: E6V

PolyPhen
>Sequence
VHLTP EKSA VTALWGKVN V DEVGGEALGR LLVVYPWTQR
FFESFGDLST PDAVMGNPKV
KAHGKKVLGA
Mutant: E6V

SNPs3D
P Yue, Z Li & J Moult (2005) JMB 353:45
>Sequence
VHLTP EKSA VTALWGKVN V DEVGGEALGR LLVVYPWTQR
FFESFGDLST PDAVMGNPKV
KAHGKKVLGA
Mutant: E6V

© Burkhard Rost
RostLab.
Misfunction/neutral
SNAP performs well

80,000 mutants w/known effects on function

Predict Features

Q2

© Yana Bromberg, 2010 Columbia University
SNAP much better for tough cases

80,000 mutants w/known effects on function

Predict Features

Q2-tough

© Yana Bromberg, 2010 Columbia University

© Burkhard Rost

Y Bromberg & B Rost 2007 NAR 35:3823-35
New directions

- in silico alanine scan
- comprehensive in silico mutagenesis
- prediction of binding hot spots

Y Bromberg & B Rost (2008) Bioinformatics 24: i207-212
**Prediction and mutagenesis**

>MC4R_HUMAN
MVNSTHRGMHTSLHLWRSSYRLHSNASESLGKGYSDG
GCYEQLFVSPEVFVTLGVLVINLVIIVAIAKKNLHSPMY
FFICSLAVADMLVSNGSEITIVITLLNSTDTDAQSFTVNI
DNDSVICSSLASICLLSIAVDRYFTIFYALQYHNIIMTVK
RGILISCICWACTVSSGIIYSDSAVIIICLITMFFMLALMAS
LYVHMFLMARLHIKRIAVLPGTGIRQGANMKGAITLTILIG
VFVVCWAPFFHLIFYISCPQNYVCVCFMSHFNLYLILIMCN
SIIDPIYALRSQELRKTFKEIICCYPLGGLCDLSSRY

R7H, S30F, E100A

<table>
<thead>
<tr>
<th>nsSNP</th>
<th>Prediction</th>
<th>Reliability</th>
<th>Exp Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>R7H</td>
<td>Neutral</td>
<td>5</td>
<td>89%</td>
</tr>
<tr>
<td>S30F</td>
<td>Non</td>
<td>4</td>
<td>82%</td>
</tr>
<tr>
<td>E100A</td>
<td>Non</td>
<td>3</td>
<td>78%</td>
</tr>
</tbody>
</table>

**ALL MUTANTS** (19 substitutions per position)

**ALL PREDICTIONS** (19 substitutions per position)

---

Y Bromberg & B Rost *Bioinformatics* 2008 24: i207-212

© Yana Bromberg
Important residues in binding sites

Y Bromberg & B Rost 2007 NAR 35:3823-35
Y Bromberg G Yachdav & B Rost 2008 Bioinformatics 15:2397-8
In silico mutagenesis

Experimental

Predicted (SNAP)

LacI repressor from E. coli
4011 mutants (12-13 substitutions/residue)
SNAP prediction accuracy: ~73%

Experimental: P Markiewicz et al. 1994 JMB 240:421-33
Y Bromberg & B Rost unpublished

© Yana Bromberg, 2010 Columbia University
Melanocortin receptor (MC4R)

Y Bromberg et al. 2009 FASEB 9:3059-69
Differential view on 2 similar receptors

Y Bromberg et al. 2009 FASEB 9:3059-69
nsSNP effect: more detail
Secondary structure (helix, strand) robust under random mutation, disorder not.

Christian Schaefer
Predict impact of nsSNP upon protein structure

Predict change from sequence alone

- Input:
  sequence: MSVKELEDKVEELLSKNYHLNEVARLKKLVGER
  mutation: E7M
- Outcome: change / no-change

Christian Schaefer
Molecular Dynamics (MD) can be very powerful
Large-scale protein flexibility analysis of single nucleotide polymorphisms using molecular dynamics simulations
Marc Offman, MKrol, B Rost, JL Sussman & AH Futerman (2011)
Validation of a molecular dynamics protein structure PREDICTION:
Comparison of an MD model with the X-ray structure of the N370S acid-β-glucosidase mutant that causes Gaucher disease.
PEDS in press.
MD2: new SNPs causing Parkinson’s Disease

Exome sequencing reveals mutations in the retromer protein VPS35 as cause for Parkinson’s disease.

Am J Hum Genet, 89: 168-75
Sets of SNPs: 
- human – Neandertal
- chimp
Homo sapiens vs H. neanderthalensis 78 nsSNPs

Sequence

Artist's rendering of Neandertal man, from Neandertal museum in Mettmann.
All rights reserved. Copyright: Johannes Krause, Max Planck Institute for Evolutionary Anthropology, Leipzig.

[Image of rendering]

SNAP

Janet Kelso
MPI Leipzig

Pääbo
MPI Leipzig

78 nsSNPs

Shaila Roessle

Yannick Mahlich

[Image of research team members]

[Graph or diagram showing SNP results]

<table>
<thead>
<tr>
<th></th>
<th>H. sapiens</th>
<th>H. neanderthalensis</th>
<th>P. troglodytes</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>41</td>
</tr>
<tr>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>5</td>
</tr>
<tr>
<td>no effect</td>
<td>no effect</td>
<td>effect</td>
<td>no effect</td>
<td>13</td>
</tr>
<tr>
<td>no effect</td>
<td>effect</td>
<td>effect</td>
<td>no effect</td>
<td>2</td>
</tr>
<tr>
<td>effect</td>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>5</td>
</tr>
<tr>
<td>effect</td>
<td>effect</td>
<td>no effect</td>
<td>effect</td>
<td>1</td>
</tr>
<tr>
<td>effect</td>
<td>effect</td>
<td>no effect</td>
<td>effect</td>
<td>2</td>
</tr>
<tr>
<td>effect</td>
<td>effect</td>
<td>effect</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Particular challenges for human
20,000 variants have an effect?
In 2050, we might answer by experiments ...
... meanwhile, we use computers

Question: do the 20,000 variants between us have an effect?
SNAP
(Screening for Non-Acceptable Point mutations)
SNAP predict effects of variants

SNAP mines wealth of experimental results by machine learning

Yana Bromberg Rutgers
Maximilian Hecht TUM

1 variant

INPUT

OUTPUT

EFFECT

NEUTRAL

original

no function
different function

© Burkhard Rost ROSTLAB. TUM
We used a structural genomics approach to obtain structural information for the SF1 family of 204 non-redundant proteins. As exemplified by the crystal structure of HiTehA (Fig. 1b), family SF1 is divided into sub-subfamilies as detailed in Supplementary Table 1, and proteins are in subfamily SF1A, closest bacterial homologues are in SF1B, and family SF3 comprises the fungal Mae1 and as exfoliative toxins; and family SF3 comprises the fungal Mae1 and Aspergillus fumigatus Ssu1 for SF3A, and Vibrio parahaemolyticus Ssu1 for SF3B.

Figure 1. Schizosaccharomyces pombe family SF1 is divided into sub-subfamilies (Supplementary Table 2), including alignment of TehA from representative Schizosaccharomyces pombe (AtSLAC1) and Haemophilus influenzae (HiTehA) and SLAC1 from Staphylococcus aureus (SF2A) and Aspergillus fumigatus (SF2B), with 100% identical sequence identity within the plant subfamily SF1A for AtSLAC1 and 95% identical within the plant subfamily SF1A for HiTehA; red diamonds mark HiTehA residues that are Variable Average Conserved (VAC) identical residues that line the central pore; and the coloured bar below residues encodes electrostatic potential at the extracellular surface. Two pertinent SF1 sequences are aligned in Fig. 1b.

Figure 2. Crystal structure of HiTehA and homology model of AtSLAC1. We used a structural genomics approach to obtain structural information for the SF1 family of 204 non-redundant proteins. As exemplified by the crystal structure of HiTehA (Fig. 1b), family SF1 is divided into sub-subfamilies as detailed in Supplementary Table 1, and proteins are in subfamily SF1A, closest bacterial homologues are in SF1B, and family SF3 comprises the fungal Mae1 and as exfoliative toxins; and family SF3 comprises the fungal Mae1 and Aspergillus fumigatus Ssu1 for SF3A, and Vibrio parahaemolyticus Ssu1 for SF3B.

Crystal structure of HiTehA and homology model of AtSLAC1. We used a structural genomics approach to obtain structural information for the SF1 family of 204 non-redundant proteins. As exemplified by the crystal structure of HiTehA (Fig. 1b), family SF1 is divided into sub-subfamilies as detailed in Supplementary Table 1, and proteins are in subfamily SF1A, closest bacterial homologues are in SF1B, and family SF3 comprises the fungal Mae1 and as exfoliative toxins; and family SF3 comprises the fungal Mae1 and Aspergillus fumigatus Ssu1 for SF3A, and Vibrio parahaemolyticus Ssu1 for SF3B.
SNAP performance

Cumulative fraction

Predicted impact of mutation (SNAP score)

neutral
effect

experimental neutral
experimental effect
SNAP performance

Cumulative fraction

-100 -75 -50 -25 0 25 50 75 100

Predicted impact of mutation (SNAP score)

neutral  effect
SNAP performance

Cumulative fraction

Predicted impact of mutation (SNAP score)
Disease variants have very strong effect
Now we can answer today:

Do our 20,000 differences matter?
Mutations between us all neutral?

![Graph showing cumulative fraction vs. predicted impact of mutation (SNAP score) for neutral and disease mutants.]

Our diversity neutral?

- Yannick Mahlich
- Dominik Achten
- Hecht
Many variants between us have effect
Many variants between us have effect
Differences between us and early human have less effect
Differences we-gorilla very neutral!
Our genetic differences matter!

Figure 1

Two pertinent SF1 sequences are aligned in Fig. 1b. We used a structural genomics approach to obtain structural information on the SF1 family of 204 non-redundant proteins. The TehA homologues are in SF1C (Fig. 1a). The other families also divide into subfamilies as detailed in Supplementary Table 1, and subfamily sequences (Supplementary Tables 1 and 2), including Ssu1 proteins and their archaeal or bacterial homologues, respectively.

Crystal structure of HiTehA and homology model of AtSLAC1. The TehA structure has been used to restrict sequence gaps to TM48, TM47, but coloured by sequence variability. The ribbon presentation was computed by the program COBALT.

ConSurf sequence variability

Surface of a homology model of AtSLAC1, Haemophilus influenzae (SpMae1) for SF3B. The TehA structure has been used to restrict sequence gaps to TM48, TM47, but coloured by sequence variability.

©2010 Macmillan Publishers Limited. All rights reserved

Our genetic differences between us and the early human disease mutants matter.
Our genetic differences matter!

- Cumulative fraction
- Predicted impact of mutation (SNAP score)
- Neutral effect
- Disease mutants
- Differences between us
- Diff (we - gorilla)
- Diff (we - early human)
Weakly effect variants may define individuality

Y Bromberg et al 2013 PNAS 110:14255-60

Scientists investigate the functional diversity of proteins
Wide range of differences, mostly unseen, among humans

doi http://www.pnas.org/content/110/35/14255

5.09.2013, Research news

No two human beings are the same. Although we all possess the same genes, our genetic code varies in many places. And since genes provide the blueprint for all proteins, these variants usually result in numerous differences in protein function. But what impact does this diversity have? Bioinformatics researchers at Rutgers University and the Technische Universität München (TUM) have investigated how protein function is affected by changes at the DNA level. Their findings bring new clarity to the wide range of variants, many of which disturb protein function but have no discernible health effect, and highlight especially the role of rare variants in differentiating individuals from their neighbors.

Barbara Wankerl - TUMnews

Y Bromberg, PC Kahn & B Rost 2013 PNAS 110:14255-60
Effects might be bad and good

Sickle cell disease (sickle cell anaemia)

- caused by single change in haemoglobin (E6V)
- mostly in malaria regions

“bad” mutation increases malaria resistance

AC Allison 1954 British Med J 1:290
What if we could curtail the bad effects?
What if we could curtail the bad effects?
personalized medicine?
One drug tailored to each?
Fewer than 40 new drugs each year

One new drug takes about **14 years** and **$2 billion** to develop.

New Drug = new NME (New Molecular Entity) approved by Federal Drug Agency (FDA), USA

data from: M Allison (2012) Nat Biotech 30, 41-49 Fig. 2
personalized health
(diagnosis/choice, food)

NOT

personalized drugs
What if we could curtail the bad effects?
Conclusion
Our genetic differences matter!

They make our individuality AND sensitive to problems.

Possibly bad for us, but we survive as a species.
Evolution risks diversity that brings change

Cumulative fraction

-100 -75 -50 -25 0 25 50 75 100

Predicted impact of mutation (SNAP score)

neutral
difference between us
diff (we - gorilla)
disease mutants
Through personalized health, the bad effects could be somehow checked!
Personalized health is harnessing the power of diversity

BEST OF BOTH WORLDS
THANK YOU

Maximilian Hecht
Yannick Mahlich
Chris Dominik Schäfer
Dominik Achten
Edda Kloppmann
Marlena Drabik
Guy Yachdav
Andrea Schafferhans
Tim Karl
Laszlo Kajan
Lothar Richter
Tatyana Goldberg

Yana Bromberg
Janet Kelso
Svante Pääbo
Avner Schlessinger
Yanay Ofran
Michal Linial
Reinhard Schneider
Chris Sander
Karima Djabali
Lena Rost

Boyang Ying
TEDx TUM

Wolfgang Herrmann
TUM

© Burkhard Rost
ROSTLAB. TUM
Personalized health: harnessing the power of diversity
Trivial step from predicting the effects of SNPs to individualized medicine?

This is going to be a challenging tough short way

connect many resources

AND look at the detail

ONLY ONE WAY: work together
Today: Intro into function

☑️ LAST WEEKs
  ● Intro into lecture

☐ THIS WEEK
  ● Tuesday: From molecular predictions to individualized medicine
  ● Introduction to protein function I

☐ NEXT WEEK
  ● Tobias Hamp: function prediction
comp bio:
continue here
this needs work!!
PUT in slide: idea of how to align