V. Predict effect of sequence variation 2

short title: cb2.snp2

lecture: Computational Biology 2 - Protein function (for Informatics) - TUM summer semester
Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- 11/29 Jonas Reeb
- 12/01 Tatyana Goldberg
- 01/19 Jana Schmidt/Carlos X (TBC)

No lecture:
- 11/01 All Saints
- 11/15 SVV Student representation
- 11/24 Thanksgiving
- 12/20-01/06 - no lecture Xmas

LAST lecture: January 19 (followed by 3 x wrap-up sessions)

Examen: February 07 time TBA

Makeup: TBC: Apr 25 & Apr 26, 2017 - lecture time
V. Predict effect of mutations
Recap
Oncogene K-Ras: single G12C mutation

structure (PDB id 4lpk-rainbow/4l8g-red):

Slide from:
Andrea Schaffermans
Sickle Cell Disease: Gain of Function

Autosomal recessive disorder
E6V in HBB causes interaction w/ F85 and L88
Formation of fibrils
Abnormally shaped red blood cells, leads to sickle cell anemia
Manifestation of disease vastly different over patients

Predrag Radivojac
© Burkhard Rost
ROSTLAB
http://gingi.uchicago.edu/hbs2.html
May be as many as 150 distinct prediction methods out there
1\textsuperscript{st} generation:
use simple small data set of variants only: SIFT, PolyPhen, SNPs3D
nsSNP effects: some *in silico* methods

**SIFT**
PC Ng & S Henikoff (2003) NAR 31:3812-14

> Sequence
VHLTEEEKSA
VTALWGKVNV
DEVGGEALGR
LLVYYPWTQR
FFESFGDLST
PDAVGNPKV
KAHGKKVLGA
Mutant: E6V

**PolyPhen**

> Sequence
VHLTEEEKSA
VTALWGKVNV
DEVGGEALGR
LLVYYPWTQR
FFESFGDLST
PDAVGNPKV
KAHGKKVLGA
Mutant: E6V

**SNPs3D**
P Yue, Z Li & J Moult (2005) JMB 353:459-63

> Sequence
VHLTEEEKSA

© Burkhard Rost
ROSTLAB. TUM
Classification of SAV-effect prediction methods

- **1st generation:**
  use simple small data set of variants only: SIFT, PolyPhen, SNPs3D

- **2nd generation:**
  accumulated larger data sets, advent of machine learning
  e.g. SNAP, SIFT2, PolyPhen-2

- **3rd generation:**
  large data sets, including diversity of information, essentially machine learning, only
SNAP: neural network

Score:
\(-100 \leq S \leq 100\)
SNAP1 input

**SNAP**
- biophysical features
- alignment profiles
- solvent accessibility (Reprof)
- secondary structure (Reprof)
- residue flexibility (PROFbval)
- AAindex + other global features
- predicted binding sites (InteractionSites, DIS)
- disordered regions
- contact potentials
- correlated mutations
- low-complexity regions

**SNAP annotated**
- SWISS-PROT annotations
- Pfam domains
- PROSITE
- SIFT predictions

**Special versions**
- no alignment
- no disease data

SNAP2 best for tough human variants

- High agreement between methods: 61%-77%
- Some predictions are more difficult than others

Classification (84k):
- Easy (54k)
- Unsolvable (6k)
- Difficult (24k)

\[ Q2 = 100 \times \frac{\text{# correct predictions}}{\text{total # of predictions}} \]
Accuracy and Reliability

V.6 SAV effect:

SNAP

• disease data
• beyond singles
Disease data
OMIM disease variants stand out

SNAP2 version not trained on OMIM:

- All methods predict many OMIM variants as effect.
- PolyPhen-2 more so than SNAP2, but at the prize of over-predicting neutral variants.
- SNAP2 predicts unknown OMIM variants with more effect than cross-validated effect training set.

OMIM disease variants stronger than training set

SNAP2 version not trained on OMIM:

- All methods predict many OMIM variants as effect
- PolyPhen-2 more so than SNAP2, but at the prize of over-predicting neutral variants

Application: Disease variants

Disease variants more effect

Mapping human disease variants to mouse:

- Fewer variants predicted as effect in mouse ortholog than in native human
- Unclear why

Application: Disease variants

For each OMIM variant alternatives with:
- conservation $\geq$ Co
- conservation $<$ Co

- more effect predicted
- Almost as much variants (not shown)
- Largely independent

$\Rightarrow$

Position of variants crucial

Beyond singles
New directions

- *in silico* alanine scan
- Comprehensive *in silico* mutagenesis
- Prediction of binding hot spots

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

Targeted Mutagenesis

Residue Scan

**comprehensive all against 19 non-native**

Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12
Prediction and mutagenesis

>MC4R_HUMAN
MVNSTHRGMHTSLHLWNRSSYRLHSNASESLGKY
SDGGGYEQFLFSPEVFVTGVLVLLENILVIVAIKKN
LHSPMYFFICSLAVADMLVSVEGSSETIVITLLNSTDT
DAQSFTVNDIVIDSVGCSSLASICSSLISIAVDRYFTIF
YALQYHNNITVKGIVIISCIWAACTEVSUGLIYSDSASA
VIVCLITMFFTMLALMASLYVMFMLMRHLHIKRKLVLPG
TGAIRQGANMKGAILTLIGVFVVFWAPFFLHLIFYIS
CPQNYPCVCFMSHFNLILIMMCNIIIDPLIYALRSQEL
RKTFKEIICYPGLGTLSSRY

nsSNP  Prediction  Reliability  Exp Accuracy
----------  ----------  -------------  -----------------
R7H  Neutral  5  89%
S30F  Non4  82%
E100A  Non3  78%

Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12

slide: © Yana Bromberg
In silico “alanine” scan

Correlation of average over 19 possible SNAP predictions per location to single residue scores

Substitute amino acids

Y Bromberg & B Rost (2008) Bioinformatics 24:i207-12
Predict binding hotspots
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
Functional residues

Mean over non-native  Conservation  Other scoring

Y Bromberg & B Rost 2007 NAR 35:3823-35
Y Bromberg G Yachdav & B Rost 2008 Bioinformatics 15:2397-8
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
Landscape of mutability

• Ultimately: Every variant (SAV) might be observed
Landscape of mutability

- Ultimately: Every variant (SAV) might be observed

hold on: Really? EVERY SAV?
Landscape of mutability

- Ultimately: Every life-compatible variant (SAV) might be observed
Landscape of mutability

M Hecht, Y Bromberg & B Rost (2013) JMB 425:3937-48
Landscape of mutability

M Hecht, Y Bromberg & B Rost (2013) JMB 425:3937-48
Ultimately: Every ‘life-compatible’ variant might be observed

SNAP2 predicts the effect for every possible variant

Which of the high-effect predictions are phenotypically important?
Analysis of deep scanning experiments
Point mutation landscape predicted experiment

SNAP2

lacl repressor

M Hecht, Y Bromberg & B Rost (2013) JMB 425:3937-48
Point mutation landscape predicted

experiment

lacl repressor

SNAP2

Binary: 1|0

how many identical?

M Hecht, Y Bromberg & B Rost (2013) JMB 425:3937-48
Point mutation landscape predicted

experiment

lacI repressor

SNAP2

Binary: 1|0: 78.2% “correct”
is that a lot?

M Hecht, Y Bromberg & B Rost (2013) JMB 425:3937-48
Point mutation landscape predicted

experiment

M Hecht, Y Bromberg & B Rost (2013) JMB 425:3937-48
Point mutation landscape predicted

(a) lacI repressor

SNAP2

experiment

Binary: 1|0: 78.2% “correct”
correlation: 0.76

M Hecht, Y Bromberg & B Rost (2013) JMB 425:3937-48
Following slides: Bachelor Thesis from Theresa Wirth

Theresa Wirth
LMU & TUM, Munich
Part 1: Experimental data

- Experimental data sets for 12 proteins
- Taken from PhD Thesis of Thomas Hopf

Deep mutational scanning
- All single amino acid mutations
- Quantitative change in protein fitness

<table>
<thead>
<tr>
<th>Position</th>
<th>WT AA</th>
<th>Mutant AA</th>
<th>Fitness</th>
<th>Estimated error in fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>M</td>
<td>A</td>
<td>0.0020</td>
<td>0.0003</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>C</td>
<td>0.0030</td>
<td>0.0005</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>D</td>
<td>0.0022</td>
<td>0.0006</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>E</td>
<td>0.0024</td>
<td>0.0004</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>F</td>
<td>0.0057</td>
<td>0.0016</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>G</td>
<td>0.0012</td>
<td>0.0002</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>H</td>
<td>0.0017</td>
<td>0.0004</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>I</td>
<td>0.3780</td>
<td>0.0491</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>K</td>
<td>0.0026</td>
<td>0.0004</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>L</td>
<td>0.3929</td>
<td>0.0511</td>
</tr>
</tbody>
</table>
First challenge: normalisation

→ Scaled and unscaled value distributions
### 12 proteins - 3 methods

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>SNAP2</th>
<th>SIFT</th>
<th>PolyPhen-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>accuracy</td>
<td>accuracy</td>
<td>accuracy</td>
</tr>
<tr>
<td>BLAT_ECOLX_2014</td>
<td>78 ± 0.48%</td>
<td>70 ± 0.53%</td>
<td>71 ± 0.52%</td>
</tr>
<tr>
<td>BLAT_ECOLX_2015</td>
<td>78 ± 0.49%</td>
<td>70 ± 0.51%</td>
<td>71 ± 0.51%</td>
</tr>
<tr>
<td>BRCA1_HUMAN_2015</td>
<td>57 ± 0.97%</td>
<td>51 ± 1.00%</td>
<td>51 ± 1.00%</td>
</tr>
<tr>
<td>DLG4_RAT_2012</td>
<td>53 ± 1.01%</td>
<td>50 ± 1.05%</td>
<td>50 ± 1.02%</td>
</tr>
<tr>
<td>FYN_HUMAN_2003</td>
<td>73 ± 5.48%</td>
<td>58 ± 6.06%</td>
<td>64 ± 5.92%</td>
</tr>
<tr>
<td>GAL4_YEAST_2015</td>
<td>62 ± 1.14%</td>
<td>57 ± 1.14%</td>
<td>56 ± 1.19%</td>
</tr>
<tr>
<td>KKA2_KLEPN_2014</td>
<td>66 ± 0.55%</td>
<td>63 ± 0.57%</td>
<td>61 ± 0.55%</td>
</tr>
<tr>
<td>MTH3_HAEEA_2015</td>
<td>77 ± 0.77%</td>
<td>73 ± 0.82%</td>
<td>74 ± 0.82%</td>
</tr>
<tr>
<td>PABP_YEAST_2013</td>
<td>58 ± 1.18%</td>
<td>58 ± 1.18%</td>
<td>68 ± 1.09%</td>
</tr>
<tr>
<td>RL401_YEAST_2013</td>
<td>51 ± 1.17%</td>
<td>50 ± 1.16%</td>
<td>54 ± 1.16%</td>
</tr>
<tr>
<td>RL401_YEAST_2014</td>
<td>50 ± 1.10%</td>
<td>50 ± 1.11%</td>
<td>53 ± 1.13%</td>
</tr>
<tr>
<td>TRY2_RAT_2009</td>
<td>27 ± 9.60%</td>
<td>47 ± 11.17%</td>
<td>40 ± 10.74%</td>
</tr>
<tr>
<td>UBE4B_MOUSE_2013</td>
<td>62 ± 1.34%</td>
<td>59 ± 1.40%</td>
<td>55 ± 1.36%</td>
</tr>
<tr>
<td>YAP1_HUMAN_2012</td>
<td>67 ± 2.04%</td>
<td>59 ± 2.14%</td>
<td>53 ± 2.11%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>68 ± 0.23%</strong></td>
<td><strong>63 ± 0.24%</strong></td>
<td><strong>63 ± 0.23%</strong></td>
</tr>
</tbody>
</table>
Tough and easy to predict correlates

Deviation from average tool behavior

\[ Z = \frac{X - \mu}{\sigma} \]
Distinguishing data sets

Accuracy based on subsets of experimental data

66% most reliable (neutral + effect)  only deleterious
Experiment & prediction somehow correlate
Different experiments agree more with each other than prediction & experiment

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Total count</th>
<th>Agreement</th>
<th>SNAP2</th>
<th>SIFT</th>
<th>PolyPhen-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAT_ECOLX_2014</td>
<td>4,783</td>
<td>95%</td>
<td>78%</td>
<td>70%</td>
<td>71%</td>
</tr>
<tr>
<td>BLAT_ECOLX_2015</td>
<td></td>
<td></td>
<td>78%</td>
<td>70%</td>
<td>71%</td>
</tr>
<tr>
<td>RL401_YEAST_2013</td>
<td>1,172</td>
<td>62%</td>
<td>51%</td>
<td>50%</td>
<td>54%</td>
</tr>
<tr>
<td>RL401_YEAST_2014</td>
<td></td>
<td></td>
<td>50%</td>
<td>50%</td>
<td>53%</td>
</tr>
</tbody>
</table>
3 methods appear similar

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>SNAP2 accuracy</th>
<th>SIFT accuracy</th>
<th>PolyPhen-2 accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>68 ± 0.23%</td>
<td>63 ± 0.24%</td>
<td>63 ± 0.23%</td>
</tr>
</tbody>
</table>
Methods rather different
V.7 SAV effect: From here
Molecular Dynamics (MD) can be very powerful
Large-scale protein flexibility analysis of single nucleotide polymorphisms using molecular dynamics simulations

© Marc Offman (TUM Munich)
Molecular dynamics of SNP in Gaucher disease

Fig. 1: Marc N Offman, M Krol, B Rost, I Silman, 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
MD2: new SNPs causing Parkinson’s Disease

Fig. 2:
A Zimprich et al.
Am J Hum Genet, 89: 168-75
Sets of SNPs:

human - Neandertal

- chimp
Homo sapiens vs. Homo neanderthalensis

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

78 nsSNPs

SNAP

Svante Pääbo
MPI Leipzig

Janet Kelso
MPI Leipzig

Shaila Roessle

Yannick Mahlich

Sequence
Data set (MPI Leipzig, S Paabo & J Kelso):
78 nsSNPs such that
• fixed in modern human (1000 genomes)
• ancestral in Neandertal & chimp

put differently:
population of modern human all do NOT have
nsSNP but Neandertal & chimp DO!
78 SNPs in 69 proteins
- only 27 (of 69) map to network (45K PPIs/10K prot)
- Surprisingly few; few pathways
- Stark contrast to other sets (e.g. human viral targets)
  -> less well-studied proteins, may be downstream effectors
V2 Sequence variance in human
Individualized medicine / personalized medicine
Map genotype to phenotype

© US Dept Health & Human Services
www.hhs.gov/ohrp/sachrp/mtgings/mtg07-08/present/koenig.html
Genotype to treatment

© Adam Simpson
http://www.adsimpson.com/pages/personalized_medicine.htm
From molecular to system


Slide: © Marc Offman (TUM Munich)
human genome completely sequenced
Genetic Code of Human Life Is Cracked by Scientists

By NICHOLAS WADE

WASHINGTON, June 26 – In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions that defines the human organism.

"Today we are learning the language in which God created life."
## Genes to proteins

<table>
<thead>
<tr>
<th>DNA / Genes</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>information, library, manual</td>
<td>machinery of life</td>
</tr>
<tr>
<td>human genome: 3 G letters</td>
<td>human proteome: ~20 k proteins</td>
</tr>
</tbody>
</table>

![DNA / Genes to Protein](image)
2015:

Human:

• 20,000 genes

• 100,000 constructs

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

- same letter
- different letter
differ by one word in this library!

4000 books
most cells in you have the same library!

4000 books

I may have 55 trillion cells *
20,000 variants have an effect?
In 2050, we might answer by experiments ...

Question: do the 20,000 variants between us have an effect?
... 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

©2010 a b c d
Cumulative performance

Cumulative fraction

Predicted impact of mutation (SNAP score)

neutral

effect
Cumulative performance

Predicted impact of mutation (SNAP score)

Cumulative fraction

neutral

effect

25%

90%

25%

98%

82%

0%
SNAP performance

Cumulative fraction

Experimental effect

Experimental neutral

Predicted impact of mutation (SNAP score)

neutral

effect

neutral

effect

© Burkhard Rost
ROSTLAB.

85/133
Protomer is colored spectrally from blue at its N terminus to red at its C terminus.

Figure 1

<table>
<thead>
<tr>
<th>A</th>
<th>SF1Aii</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>SF1Aia</td>
</tr>
<tr>
<td>C</td>
<td>SF1Bi</td>
</tr>
<tr>
<td>D</td>
<td>SF1Bii</td>
</tr>
<tr>
<td>E</td>
<td>SF1Biii</td>
</tr>
<tr>
<td>F</td>
<td>SF1Cii</td>
</tr>
<tr>
<td>G</td>
<td>SF1Ciii</td>
</tr>
</tbody>
</table>

Electronegative and electropositive potential are colored in degrees of red and blue saturation, respectively.

Sequence analysis for the SLAC1 superfamily. AtSLAC1 from Staphylococcus aureus, SF1Aii from representative Staphylococcus aureus, SF1Aia from A. thaliana, SF1Ai from Arabidopsis thaliana, SF1B from Vibrio parahaemolyticus, SF1Bii from Aspergillus fumigatus, SF1Biii from Haemophilus influenzae, SF2A from Vibrio parahaemolyticus, SF2B from Aspergillus fumigatus, SF2C from Haemophilus influenzae, SF2D from Aspergillus fumigatus, SF2E from Vibrio parahaemolyticus, SF3A from A. thaliana, SF3B from Arabidopsis thaliana, SF3C from Arabidopsis thaliana, SF3D from Vibrio parahaemolyticus, SF3E from Haemophilus influenzae, Mae1 for SF3B.

ConSurf sequence variability for SF1Aii, SF1Aia, SF1Ai, SF1B, SF1Bii, SF1Biii, SF1Cii, SF1Ciii, SF2A, SF2B, SF2C, SF2D, SF2E, SF3A, SF3B, SF3C, SF3D, SF3E, Aspergillus fumigatus, Haemophilus influenzae, Staphylococcus aureus, Vibrio parahaemolyticus, A. thaliana.

Figure 2

<table>
<thead>
<tr>
<th>A</th>
<th>SF1Ai</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>SF1Bi</td>
</tr>
<tr>
<td>C</td>
<td>SF1Bii</td>
</tr>
<tr>
<td>D</td>
<td>SF1Biii</td>
</tr>
<tr>
<td>E</td>
<td>SF1Cii</td>
</tr>
<tr>
<td>F</td>
<td>SF1Ciii</td>
</tr>
</tbody>
</table>

Electrostatic potential for the SF1 family of 204 non-redundant proteins. SF1Aii, SF1Aia, SF1Ai, SF1B, SF1Bii, SF1Biii, SF1Cii, SF1Ciii, SF2A, SF2B, SF2C, SF2D, SF2E, SF3A, SF3B, SF3C, SF3D, SF3E, Aspergillus fumigatus, Haemophilus influenzae, Staphylococcus aureus, Vibrio parahaemolyticus, A. thaliana.
**SNAP performance**

![Graph showing SNAP performance](image)

- Cumulative fraction vs. Predicted impact of mutation (SNAP score)
- Green curve: Experimental neutral
- Red curve: Experimental effect
- Blue curve: Power curve
- Pink curves: Experiment vs. neutral and experiment vs. effect

**Key Points**
- The graph illustrates the SNAP performance with different curves representing various impact categories.
- The x-axis represents the predicted impact of mutation, and the y-axis shows the cumulative fraction.
Disease variants have very strong effect
Now we can answer today:
Do our 20,000 differences matter?
Mutations between us all neutral?

The diagram shows a cumulative distribution plot with the x-axis labeled as "Predicted impact of mutation (SNAP score)" and the y-axis labeled as "Cumulative fraction". The plot includes curves for experimental neutral, disease mutants, and neutral effect. The curves are color-coded: green for neutral, orange for disease mutants, and red for neutral effect. The x-axis ranges from -100 to 100, with increments of 25.

In the bottom left corner, there is a ribbon diagram of an HiTehA protomer viewed from outside the membrane, with the ribbon colored spectrally as in Fig. 1b. The diagram also includes a family tree showing the division of the SLAC1 superfamily into subfamilies as detailed in Supplementary Table 1, and the family SF1 is divided into sub-subfamilies (Supplementary Table 2).

The diagram includes references to various proteins and bacteria, such as Staphylococcus aureus, Vibrio parahaemolyticus, Haemophilus influenzae, and Arabidopsis thaliana. The Proteobacteria and Bacteroidetes are mentioned as the closest bacterial homologues, with the archaeal homologues in SF3A and SF2A for SF1A and SF1B, respectively.

At the bottom right corner, there are images of proteins and structures, including a ribbon for the HiTehA protomer, a molecule with a green label, and a molecule with a blue label.

In the bottom left corner, there is text reading "Mutations between us all neutral?"
Many variants between us have effect

Cumulative fraction

Predicted impact of mutation (SNAP score)
Many variants between us have an effect
Where will variants to early hominids fall (e.g. Homo neanderthalensis)?
Many variants between us have an effect
Differences between us and early human have less effect.

Cumulative fraction

Predicted impact of mutation (SNAP score)
Where will variants to gorilla / chimpanzee (Gorilla gorilla/Pan troglodytes) fall?
Differences we-gorilla very neutral!

Predicted impact of mutation (SNAP score)
Differences *we-mouse* and *we-fly* even more neutral
Our genetic differences matter!

![Graph showing cumulative fraction vs. predicted impact of mutation (SNAP score).]

- **Cumulative fraction**
  - 100%
  - 75%
  - 50%
  - 25%
  - 0%

- **Predicted impact of mutation (SNAP score)**
  - Neutral
  - Effect

- **Curves**
  - Differences between us (we-early human)
  - Differences between us (we-gorilla)
  - Disease mutants

- **Markers**
  - 25%
  - 40%
Our genetic differences matter!

Cumulative fraction

Predicted impact of mutation (SNAP score)  Rost

neutral  effect

difference between us  diff (we - gorilla)
disease mutants  diff (we - early human)
... but those that matter do not happen often right?
Common variants have more effect.
Common variants have MORE effect
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

Proteins such as the interleukin-2 molecule depicted here can occur in

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
Weak-effect variants may define individuality

doi http://www.pnas.org/content/110/35/14255
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?
Sickle cell anemia: single amino acid change


AC Allison 1954 British Med J 1:290

Autosomal recessive disorder

E6V in HBB causes interaction with F85 and L88

Formation of fibrils

Abnormally shaped red blood cells, leads to sickle cell anemia

Manifestation of disease vastly different over patients

"bad" mutation increases malaria resistance

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 doi:10.1038/nbt.2083 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.
Through personalized health, the bad effects could be somehow checked!
Personalized health is harnessing the power of diversity

BEST OF BOTH WORLDS
Molecular understanding of disease
Today

color correlates with repair hours
Protein function in precision medicine: deep understanding with machine learning

B Rost, P Radivojac & Y Bromberg (2016)
FEBS Lett 590:2327-41

Machine Learning

“bad” mutation increases malaria resistance

4hhb

E6V

2hbs

Molecular function
Biological process
Cellular component

© Burkhard Rost

ROSTLAB.
Personalized health: harnessing the power of diversity
THANK YOU

Maximilian Hecht
Yannic Mahlic
Chris Dominik Achter
Edda Kloppmann
Marlena Drabik
Guy Yachdav
Andrea Schafferhans
Tim Karl
Laszlo Kajan
Lothar Richter
Tatyana Goldberg

Maximilian Hecht
Yannic Mahlic
Chris Dominik Achter
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
Afterthought
TUM course: JavaScript + DataMining + Machine Learning

Big Idea & BIG pull:
couple with introduction to computational biology
TUM course: JavaScript + Data Mining + Machine Learning

Big Idea & BIG pull:
couple with introduction to computational biology

© Wiley Miller
TUM course: JavaScript + DataMining + Machine Learning

New idea: something else
got.show: prediction of odds to survive

TUM course
JavaScript + DataMining

Guy Yachdav  Tatyana Goldberg  Christian Dallago
TUM electro car in < 15 months

Visio.M - new electrical

160km range
at 120kph
on 15kW engine
got.show: prediction of odds to survive

TUM course
JavaScript + DataMining

Guy Yachdav  Tatyana Goldberg  Christian Dallago

> 10 TV shows
> 5 radio shows
> 600 printed newspapers
> 1.2 billion potential readers
Thanks & Bye