UNIVERSITÄT DES SAARLANDES

Precision diagnostics

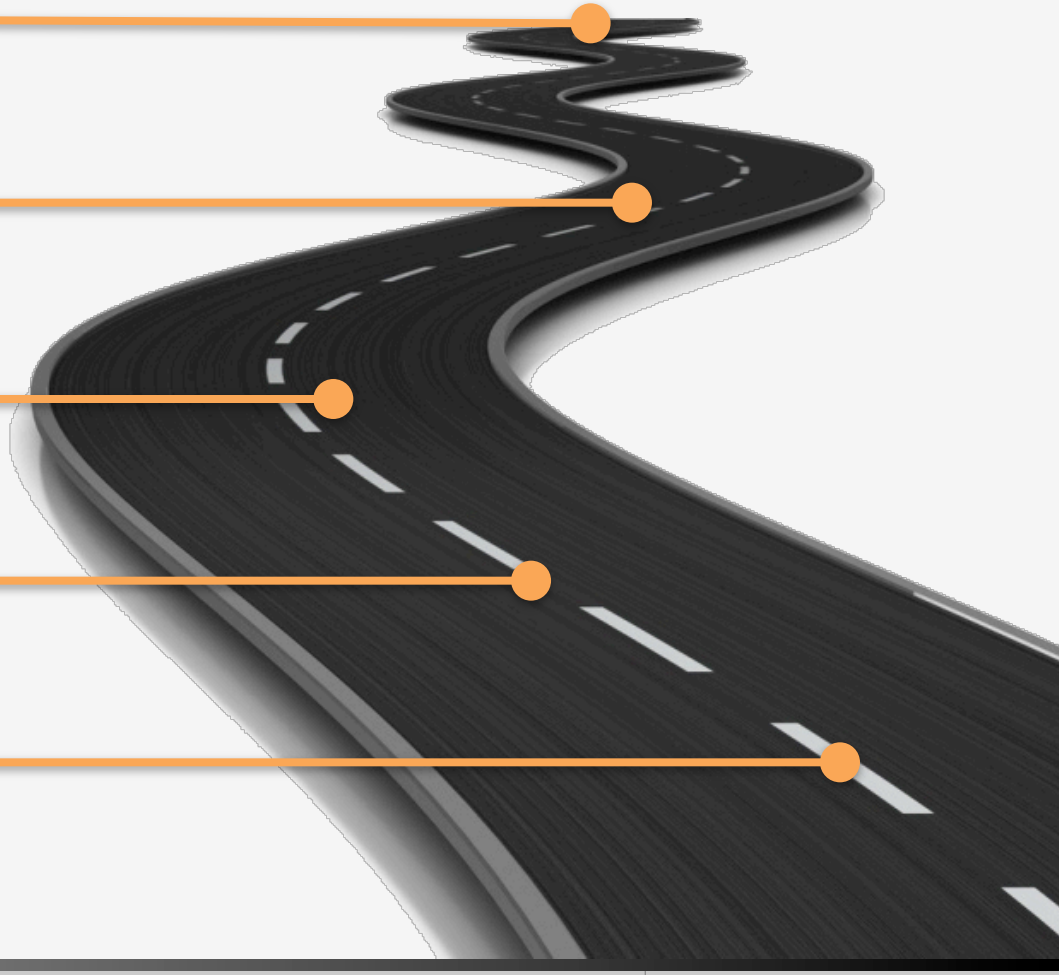
the importance of bioinformatics for Next Generation Sequencing

Saarbrücken, 11th of November, 2013

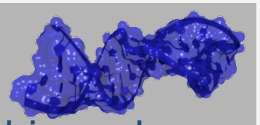
Professor Dr. Andreas Keller

Chair for Clinical Bioinformatics
Saarland University
University Hospital

- Introduction
- Genome Sequencing
- Exome Sequencing
- Gene Panel Sequencing
- Other applications

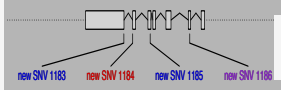


Non-Invasive Biomarkers

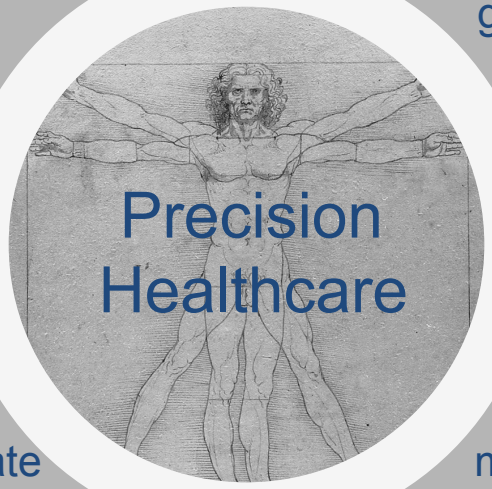


Detection of miRNA and protein biomarker patterns from human blood or serum samples using microarrays, NGS, qRT-PCR & mass spectrometry. Biostatistical evaluation & validation of the complex profiles.

Genetic Testing by NGS



Whole genome, exome or gene panel sequencing of DNA in order to detect genetic causes for human diseases. Understanding the effect of the respective genetic variants for different disease phenotypes.

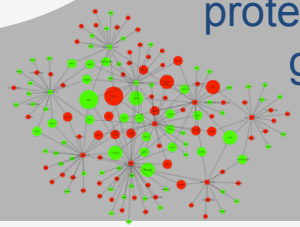


Bacterial Resistance

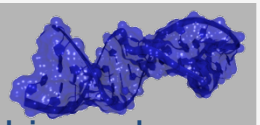
Understanding the genetic cause of bacterial resistance and correlate the bacterial resistance to classical culture based tests in order to derive the minimal inhibitory concentration and best therapy with anti bacterial agents.

Systems Biology

Model and understand how DNA, mRNA, microRNA, methylation and proteins mutually interact in order generate a holistic and multi-scale molecular representation of human pathologies.

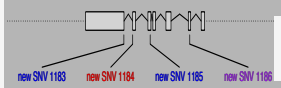


Non-Invasive Biomarkers

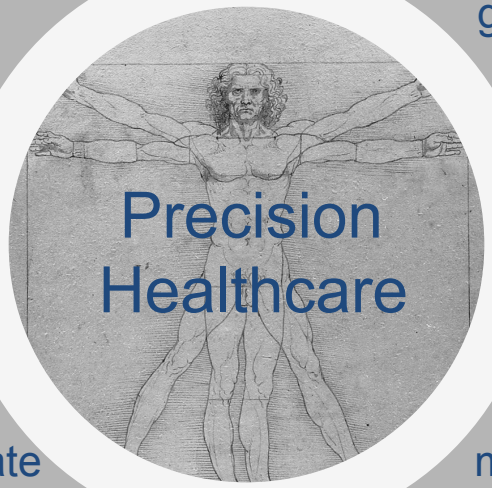


Dete…otein biomarker patt… or serum sam… NGS, qRT…etry. Bios…a- lidat…les.

Genetic Testing by NGS

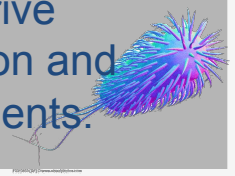


Whole g… panel sequen… detect genetic… eases. U… of the r… nts for …types.



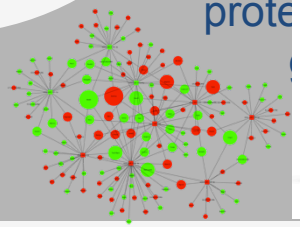
Bacterial Resistance

Und…cause of b…correlate the …lassical cultu… to derive the …entration and best…rial agents.



Systems Biology

Mod…DNA, mRNA… and prote… order …multi- …enta- …gies.



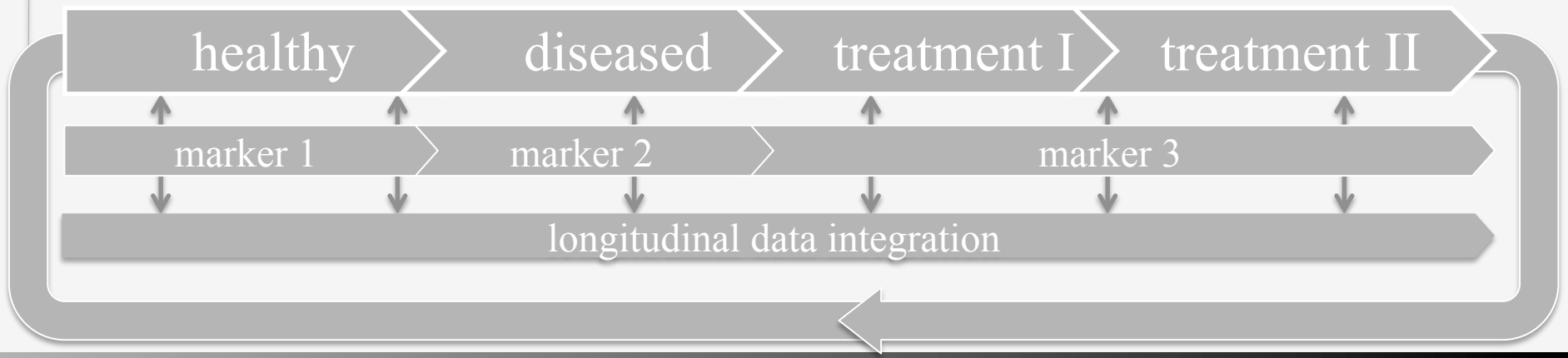
The vision:

- Using advanced informatics and biomarkers in order to…
  - deliver the right treatment – to the right patient – at the right time
  … for improving patients outcome

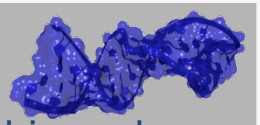
cDX - strongly growing from a weak basis

- Currently, below 5% of drugs on the market have cDX
- 35% of late development pipelines (phase IIb –IV) relying on biomarkers
- 58% of preclinical trials relying on biomarker data

An approach

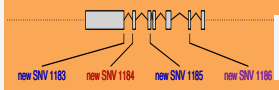


Non-Invasive Biomarkers

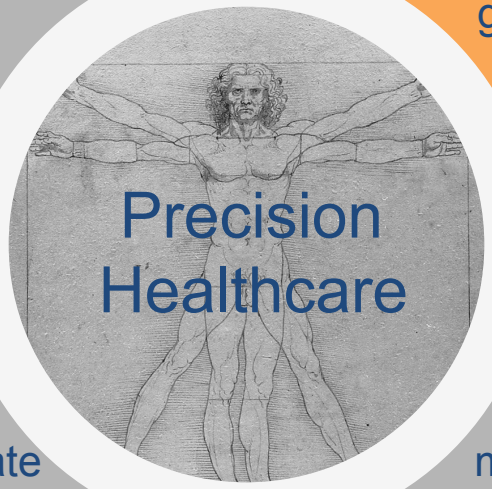


Detection of miRNA and protein biomarker patterns from human blood or serum samples using microarrays, NGS, qRT-PCR & mass spectrometry. Biostatistical evaluation & validation of the complex profiles.

Genetic Testing by NGS



Whole genome, exome or gene panel sequencing of DNA in order to detect genetic causes for human diseases. Understanding the effect of the respective genetic variants for different disease phenotypes.

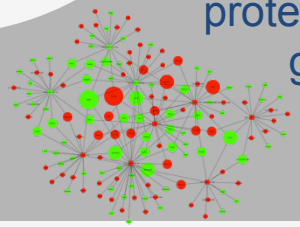


Bacterial Resistance

Understanding the genetic cause of bacterial resistance and correlate the bacterial resistance to classical culture based tests in order to derive the minimal inhibitory concentration and best therapy with anti bacterial agents.

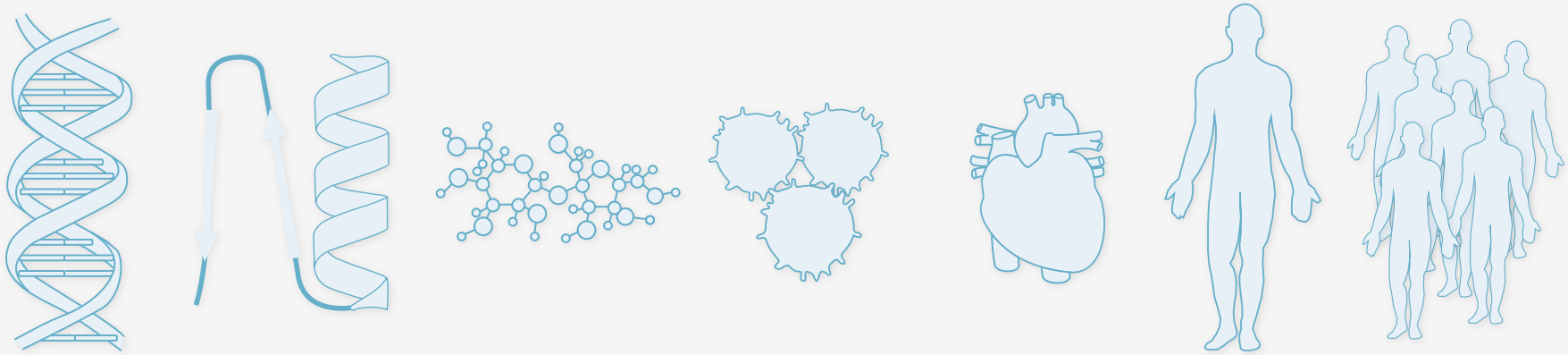
Systems Biology

Model and understand how DNA, mRNA, microRNA, methylation and proteins mutually interact in order generate a holistic and multi-scale molecular representation of human pathologies.

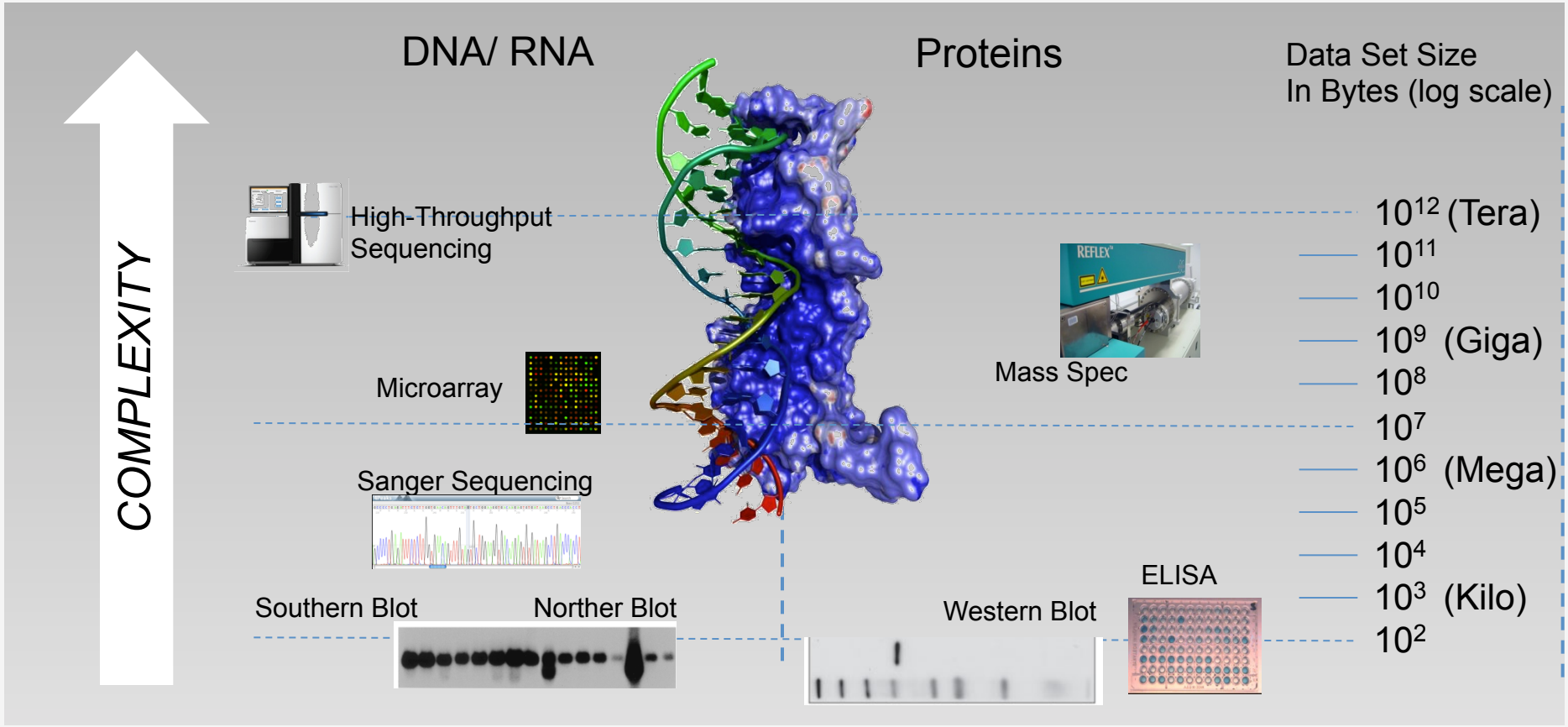


Multi Scales – from molecules to populations

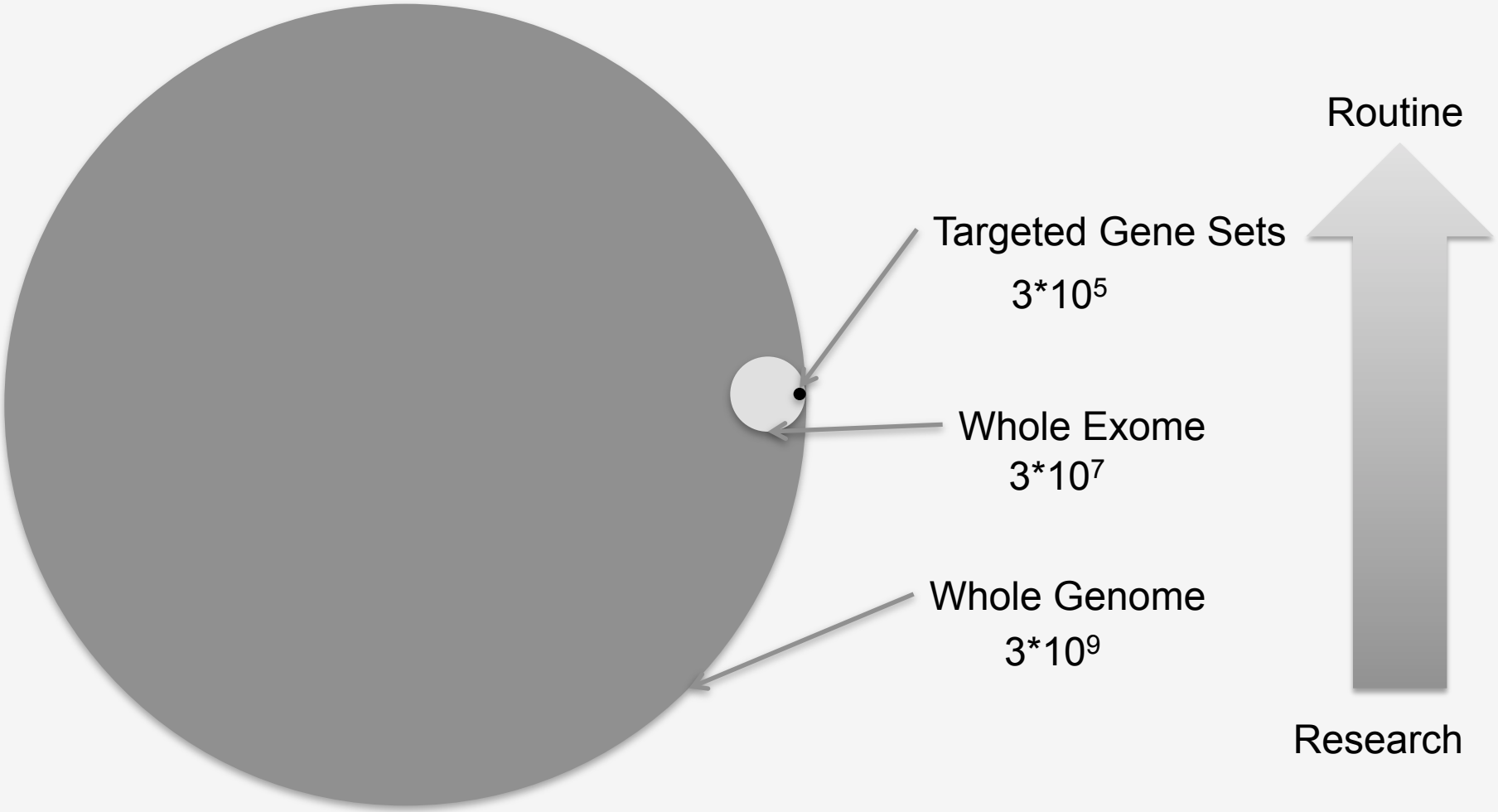
DNA, RNA | Proteins | Metabolites | Cells / Tissue | Organs | Body | Society



25.000	100.000	2.400	7 trillion	9	1	7 billion
Genes	Proteins	Compounds	Human cells	Organ-systems	Human being	People
			4 Basic types of tissue	29 Organs		



In molecular diagnostics there is a clear shift towards exponentially increasing complexity – the data can not be interpreted without IT support – Bioinformatics becomes clinically relevant


Routine
Targeted Gene Sets
3*10^5
Whole Exome
3*10^7
Whole Genome
3*10^9
Research

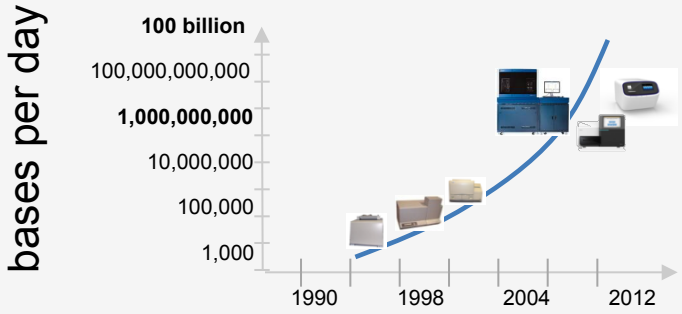
Sanger

The 10 year Human Genome Project sequenced the first human reference genome the cost of roughly $3 billion

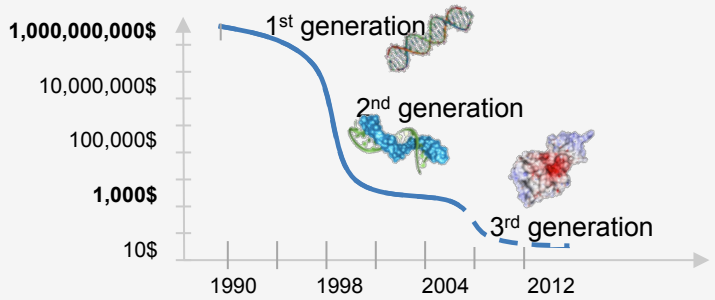
HiSeq

Today, a genome is sequenced for < $5,000 in less than 2 weeks on a single sequencing machine

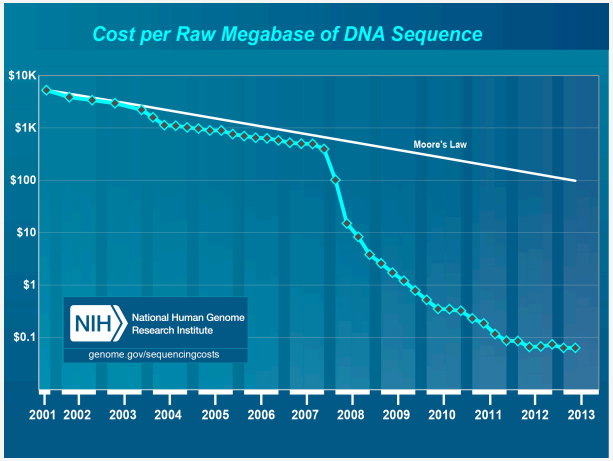
Throughput

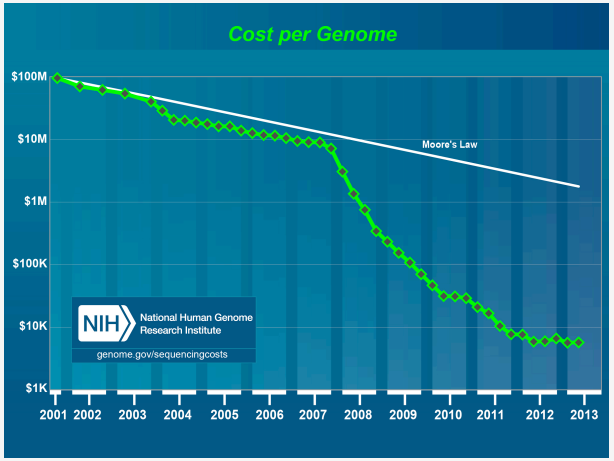


Cost per genome



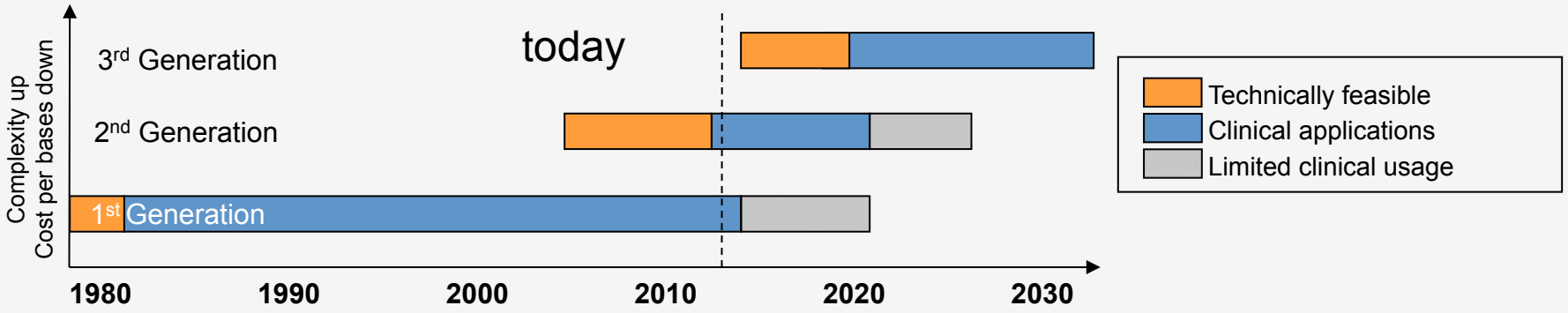
NHGRI - analysis





http://www.genome.gov/sequencingcosts/

Three generations of sequencing



First generation

- Classical sequencing approaches that are purely serial – Most relevant examples: Maxam-Gilbert Sequencing and Sanger Sequencing

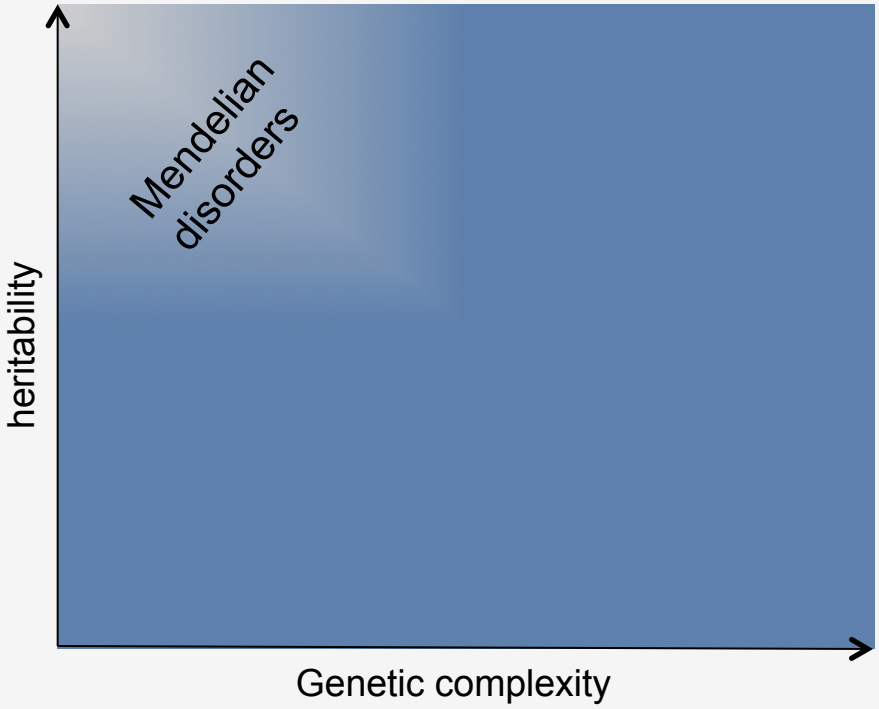
Second generation

- High-throughput and parallel sequencing approaches that do not have single cell / genome resolution – Illumina GA, HiSeq, ABI SOLiD, IonTorrent

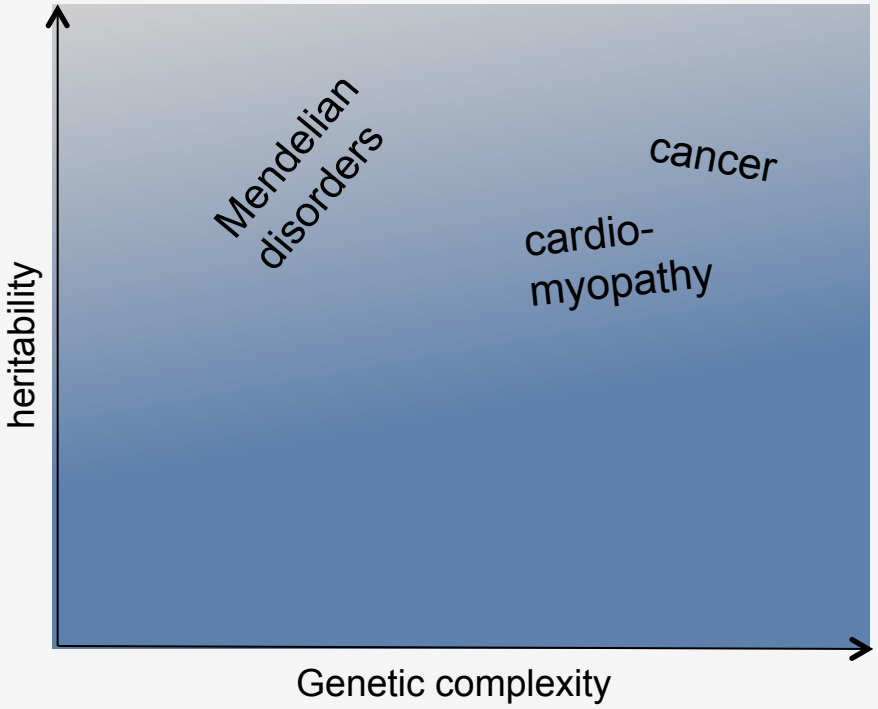
Third generation

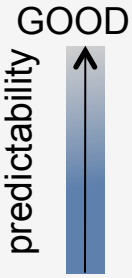
- Nanopore based sequencing approaches and single cell / genome resolution approaches – oxford Nanopores, PacBio

Sanger Sequencing



NGS





While many tests for monogenetic disorders (Mendelian Disorders) as cystic fibrosis can be carried out using Sanger sequencing for complex genetic disorders as cardiomyopathies or especially cancer Sanger sequencing lacks throughput. In addition, already for a single gene NGS is less expensive and equally accurate as Sanger sequencing.

genetic com

37 technologies…

Illumina
Ion Torrent
Roche-454
AB-SOLiD
Helicos
Pacific Bio
OxfordNanopore
Polonator
CGI
Intelligent Bio
Genapsys
Electronic Biosci
Nabsys
IBM-Roche
NobleGen
Genia
LightSpeed
GnuBio
Bionanomatrix
Halcyon
ZS Genetics
Genizon BioSci
LaserGen
Visigen/Starlight
GE Global
Stratos Genomics
Reveo
Base4innovation
Li-Cor
U.S. Genomics
Mobious Genomics
Nanophotonics Biosci
Network Biosystems
SeiraD
Affymetrix
Population Gen Tech
AQI Sciences

… with very heterogeneous performance metrics ….

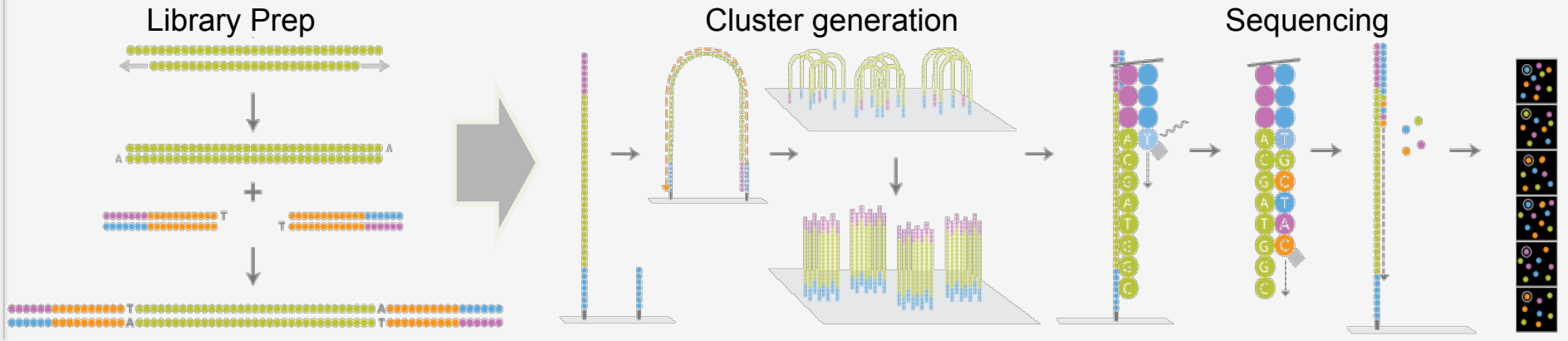
Property	Minimum	Maximum
Read length	100	1,000
# reads	200,000	6,000,000,000
Output in GB	0.5	600
Accuracy	95%	99.95%
Price per GB	$ 10	$ 4,000
Price per device	$ 100,000	$ 1,000,000
GB per day	0.1	50

… and the biggest problem becomes bioinformatics

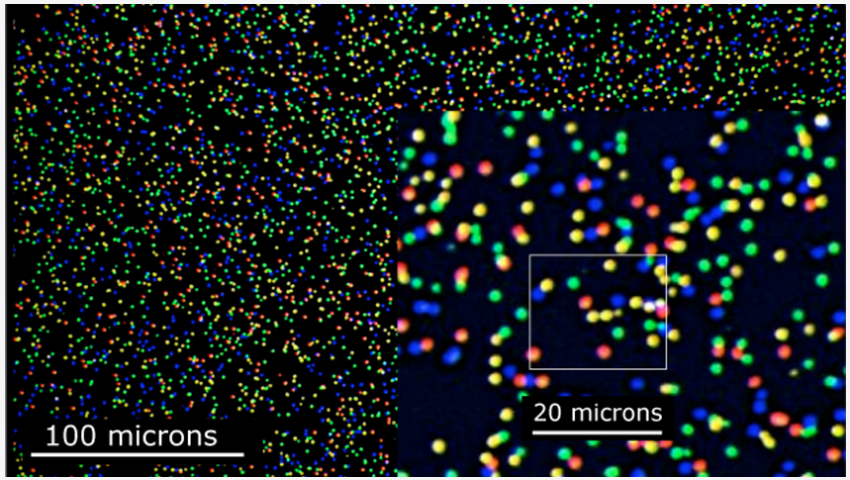
One human genome consists of ….
… up to 500 GB raw data
… 3 billion reads of 100 bases

And has to be interpreted by clinicians

Workflow



Image



Size comparison

- 50 µm — typical length of a human liver cell, an average-sized body cell
- 78 µm — width of a pixel on the display of the iPhone 4 (Retina Display)
- 90 µm — paper thickness on average

> Steps of NGS bioinformatics analysis

NGS analytics pipeline

NGS
0.3-8 days

1 – 500 GB raw data (fastq)

Primary Analysis
0.5 – 8 days

Alignment of reads to target genome

Variant calling

Detection of genomic aberrations

1 – 500 GB alignment file (bam)
1 – 500 MB variant file (vcf)

Downstream analysis
0 – 4 weeks

Detect the single disease causing variant

Understand influence on pathways

Understand influence on protein structure and therapy

Segregation analysis

Carry out targeted therapy descriptions

50 kb clinical report

Solutions

Commercial available

PROTOTYPES

Knome

INGENUITY SYSTEMS

genomatix

CLC bio
Accelerating Scientific Research

SAP

Basics

- Given an alphabet Σ = {A,C,T,G}
- Given a set S of k sequences S = {s_1,..,s_k} on the alphabet Σ, a sequence alignment is a set A = {a_1,..,a_k} of sequences on the alphabet Σ' = {A,C,T,G,-} such that
  - All sequences of A are of the same length
  - After removal of {-}, $a_i = s_i$ for all I
  - In all columns at least one character of Σ has to be

Original algorithms

- Global alignment – Needleman-Wunsch Dynamic Programming
- Local alignment – Smith-Waterman Dynamic Programming
- Heuristics such as BLAST

➔ Not suited for NGS because of runtime constrains

Mapping of reads

- Given several billions (!) of short reads (length approx. 200 bases) find the best hit of the read in the human reference genome of 3 billion bases

NGS Mapping / Alignments

- Given a set Q of k reads Q = $\{q_1,..,q_k\}$ and a reference X on an alphabet Σ = {A,C,T,G}. Find the best hit of q_i in X for all i.

NGS Mapping / Alignments

- Given a set Q of k reads Q = $\{q_1,..,q_k\}$ and a reference X on an alphabet Σ = {A,C,T,G}. Find the best hit of q_i in X for all i.

Approaches (1)

Hash Table based approaches

Building a hash of reads and scanning the genome

Eland (Cox, 2007)
RMAP (Smith, 2008)
MAQ (Li, 2008)
ZOOM (Lin, 2008)
SeqMap (Jiang 2008)
CloudBurst (Schatz, 2009)
SHRiMP (2009)

Flexible memory usage

Good runtime for large sets of reads but overhead for small sets

Usually high accuracy

Frequently problems in gap handling

Hard to be parallelized

NGS Alignments

- Given a set Q of k reads Q = $\{q_1,..,q_k\}$ and a reference X on an alphabet Σ = {A,C,T,G}. Find the best hit of q_i in X for all i.

Approaches (2)

Hash Table based approaches

Building a hash of the genome

SOAPv1 (Li, 2008)
PASS (Campagna, 2009)
MOM (Eaves, 2009)
ProbeMatch (Jung Kim, 2009),
NovoAlign, ReSEQ, Mosaik, BFAST

Large memory requirement for indexing the human genome

Easy to be parallelized, faster

Speed is determined by error rate

NGS Alignments

- Given a set Q of k reads Q = $\{q_1,..,q_k\}$ and a reference X on an alphabet Σ = {A,C,T,G}. Find the best hit of q_i in X for all i.

Approaches (3)

String matching using Burrows-Wheeler Transform

SOAPv2
Bowtie (Langmead, 2009)
BWA (Li, 2009)

Very fast at acceptable accuracy

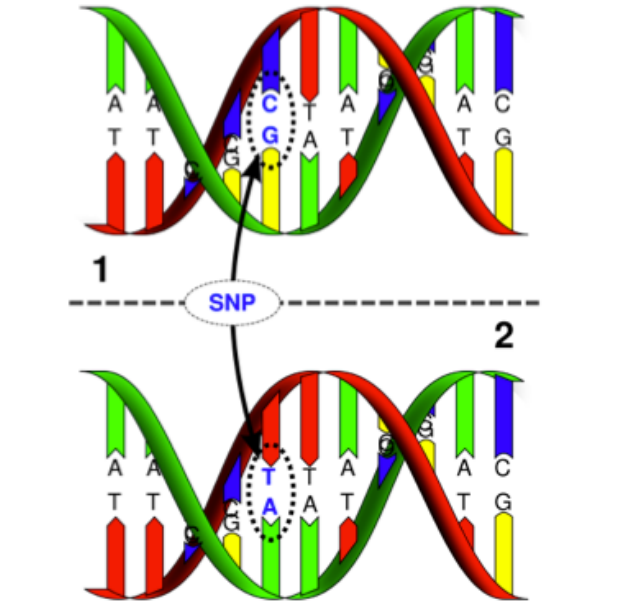
Humans are not equal

- On average humans differ approximately at every 1000th bp from the reference genome (depending on degree of relatedness, this may vary however substantially, extreme: identical twins)

SNP & SNV

- In the case of difference from the reference genome the most common alteration are SNPs (single nucleotide polymorphisms) and SNVs (single nucleotide variants)
- Other differences are short or large insertions or deletions (INDELs) or larger genomic aberrations

- Bioinformatics challenge: find the true variants and differentiate them from sequencing errors

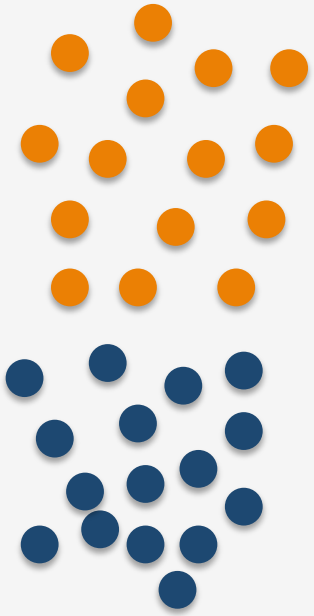


http://www.science.marshall.edu/murraye/341/Images/416px-Dna-SNP_svg.png

General issue

- We have many more DNA molecules than reads in our sequencing result.

Reference: ● Variant: ●



Original sample

Sequencing result

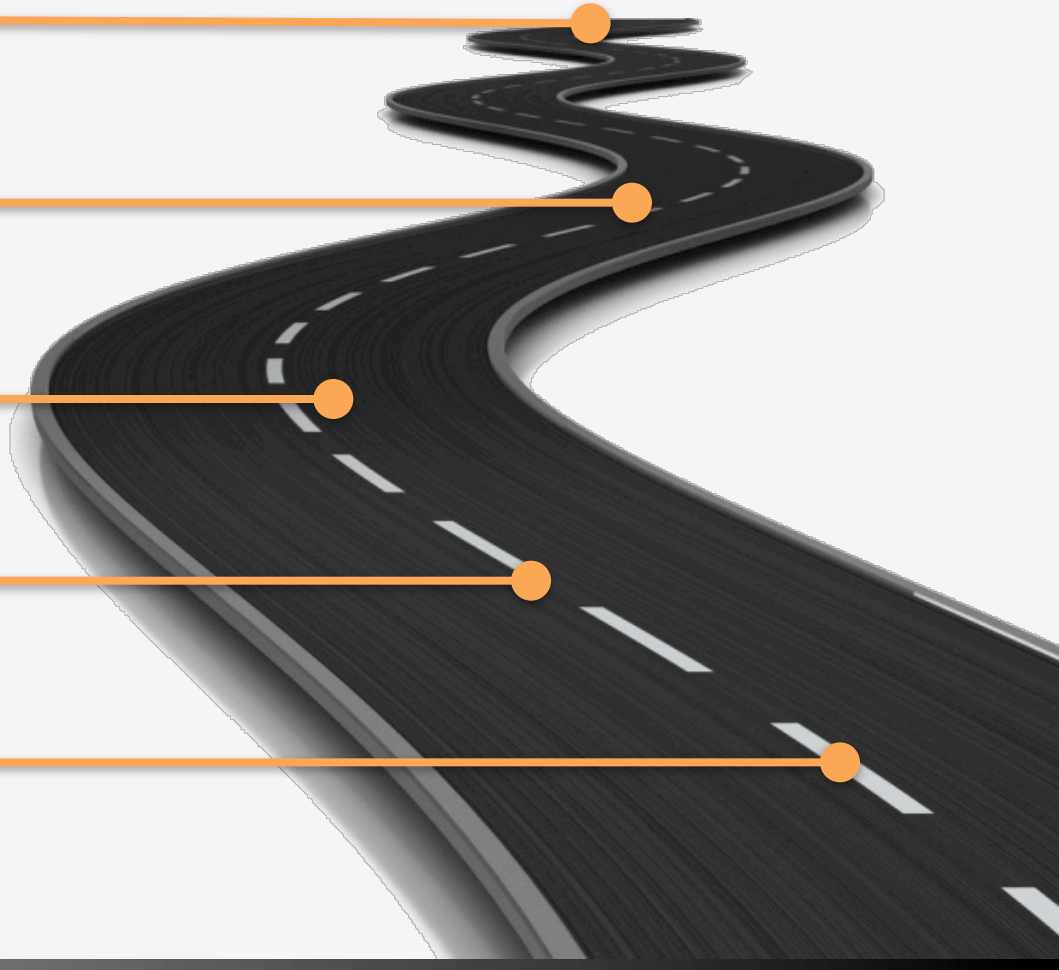
General issue

- We have many more DNA molecules than reads in our sequencing result.
- All sequencers besides single molecule NGS include a PCR step

Reference: ● Variant: ●

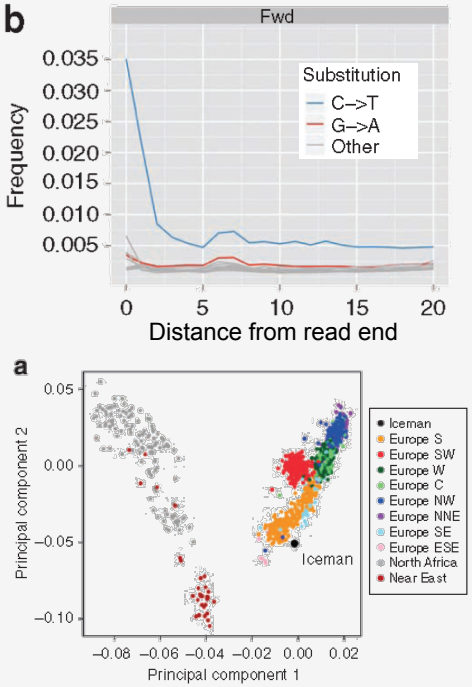
Original sample PCR result Sequencing result

- Introduction
- Genome Sequencing
- Exome Sequencing
- Gene Panel Sequencing
- Other applications



From our first whole genome sequencing project starting in 2009 …

- ABI SOLiD 4
- 3 full slides have been sequenced
- 3 billion paired end reads of read length 50 bases
- 96% coverage of the 3.2 billion bases
- Average coverage after removing duplicates was 7.6 fold
- Data evaluation took 12 months
- Sequencing costs were around 40,000 €




ARTICLE

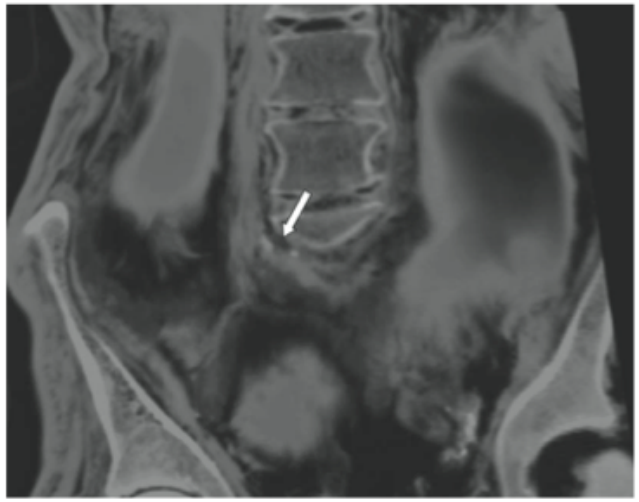
New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing

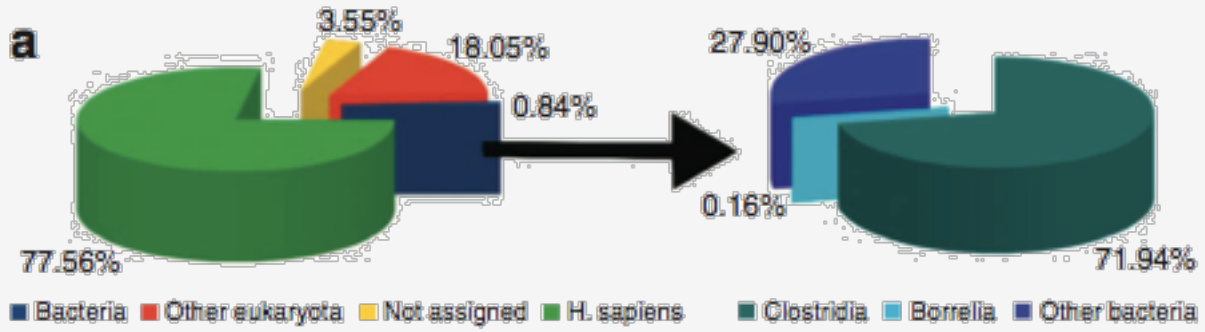
Keller et al. Nature Com. 2012

dbSNP # (b126)	Association	Forward primer 5′–3′	Reverse primer 5′–3′	Fragment size (bp)	AT (°C)	Independent PCR replication results	NGS data coverage	HapMap frequency of Iceman's genotype (sample size)[a]	
								CEU	TSI
rs10757274	Coronary artery disease	CCCCCGTG GGTCAAA TCTAAG	AGAATTCCC TACCCCTAT CTCCTATCT	82	55	nsa	8G, 1A	NA	NA
rs2383206	Coronary artery disease	TACTATC CTGGTTGC CCCTTCTGTC	GGTTCAGGA TTCAGGCCA TCTTG	78	55	G/G	8G	G/G=0.246 (130)	NA
rs5351	Atherosclerosis	TCATCCCTA TAGTTTTAC	ATGGCCAAT GGCAAGCAGA	74	55	C/T	20T, 14C	C/T=0.416 (226)	C/T=0.567 (194)

Cardiologically relevant variants

CT image of abdomen and coronal reconstruction





- Most abundant bacteria: Clostridia (72% of bacterial reads)
- 0.16% of the total bacte- rial hits assigned to sequences of the pathogen B. burgdorferi
- Around 60% of the genome covered
- But: cross-organism mapping may cause false positive hits (see also Ames et al., 2013)

… to standardized High Throughput Whole Genome Sequencing

- ABI SOLiD 4
- 3 full slides have been sequenced
- 3 billion paired end reads of read length 50bases
- 96% coverage of the 3.2 billion bases
- Average coverage after removing duplicates was 7.6 fold
- Data evaluation took 12 months
- Sequencing costs were around 40,000 €

6 fold

8 fold

- Illumina HiSeq
- 4 lanes per genome
- 1.5 billion paired end reads of read length 200 bases
- >98% coverage of the 3.2 billion bases
- Average coverage after removing duplicates of up to 45-fold
- Data evaluation takes 2-4 days
- Sequencing costs are around 5,000 €

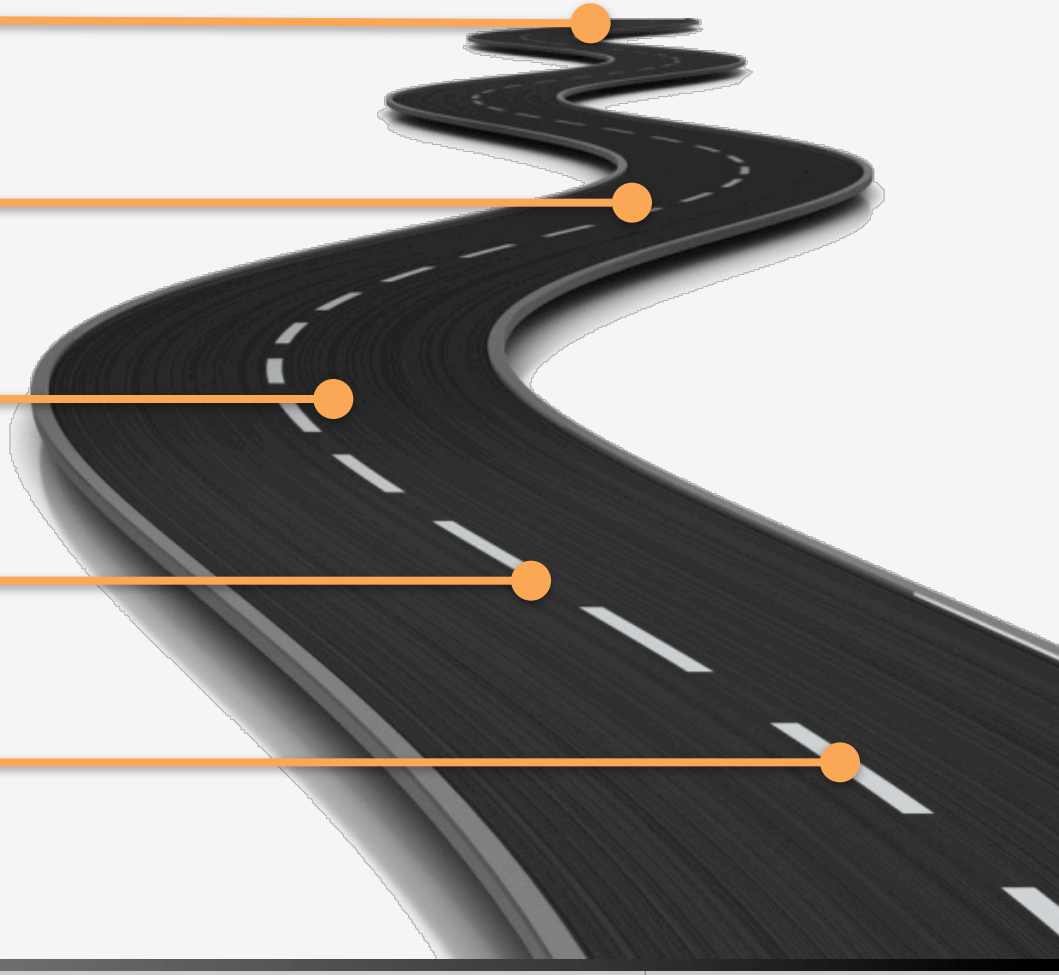
Introduction

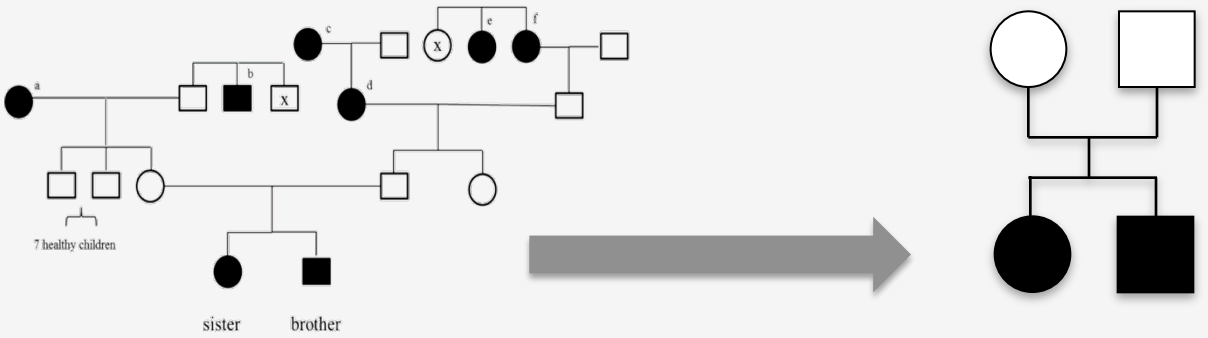
Genome Sequencing

Exome Sequencing

Gene Panel Sequencing

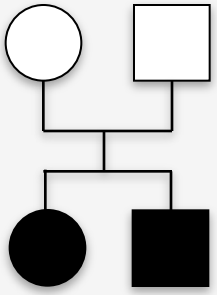
Other applications





Measuring 6 Exomes: Parents Leukocyte Genomes
Children Leukocyte Genomes
Children Tumor Genomes

% covered ≥1x	% covered ≥8x	% covered ≥20x	mean coverage	Gb of coverage
96.02	89.84	79.82	57.00	3.54
95.86	88.95	75.98	43.63	2.71
96.10	90.97	82.53	57.96	3.60
95.97	90.72	82.23	59.95	3.72
95.57	89.28	78.41	50.60	3.14
95.56	89.59	79.10	51.82	3.22
95.85	**89.89**	**79.68**	**53.49**	**3.32**

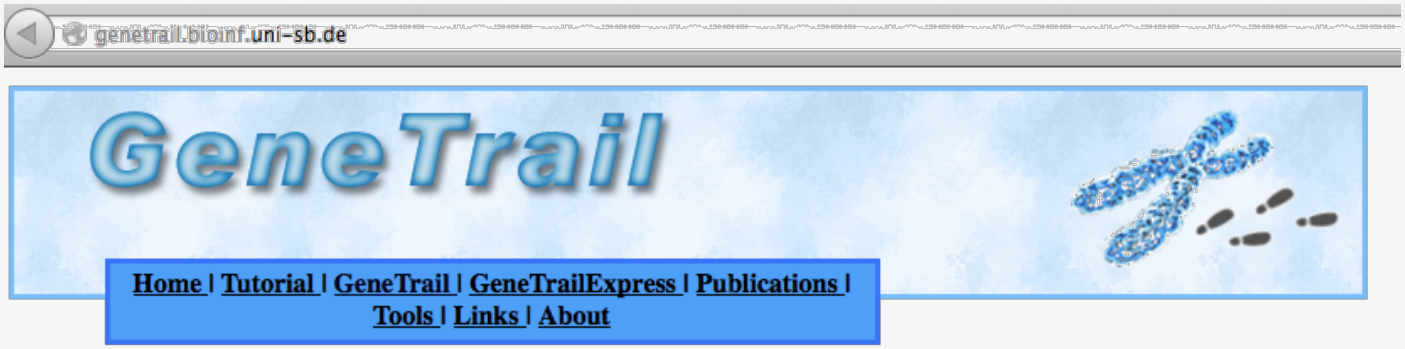


Measuring 6 Exomes: Parents Leukocyte Genomes
Children Leukocyte Genomes
Children Tumor Genomes

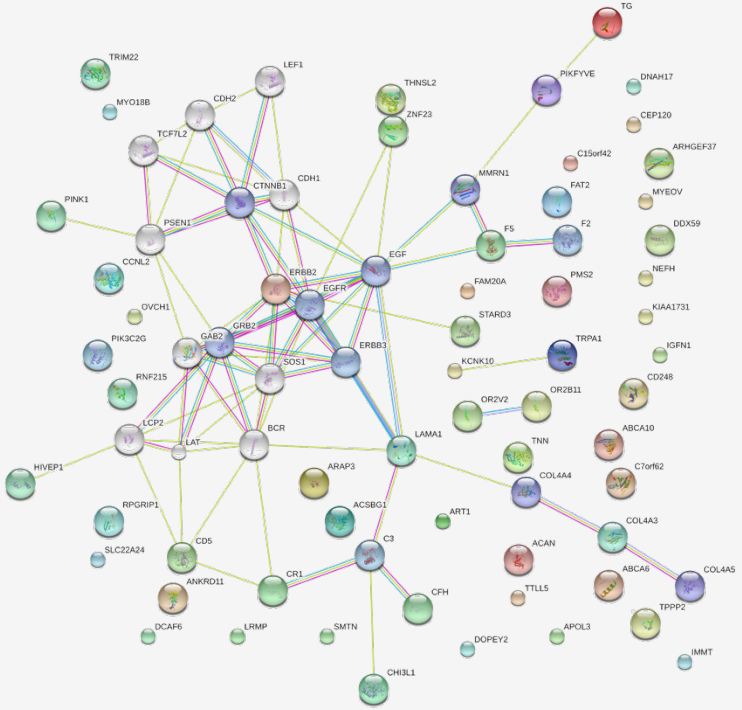
Filtering: children leukos

	A0463	**A0465**
	son, leuko	daughter, leuko
unfiltered SNPs	105,189	98,371
↓	↓	↓
not in healthy controls	26,388	22,231
↓	↓	↓
1000G<1%	8,134	7,340
↓	↓	↓
non-synonymous*	**585**	**584**
↓		
	314	

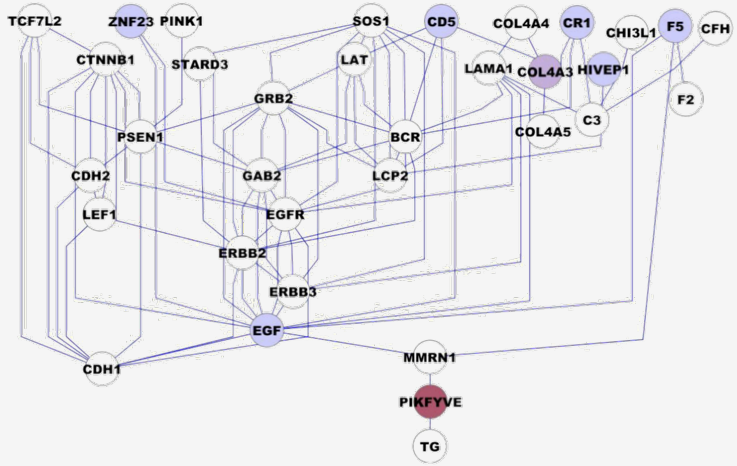
1.


genetrail.bioinf.uni-sb.de
GeneTrail
Home | Tutorial | GeneTrail | GeneTrailExpress | Publications |
Tools | Links | About

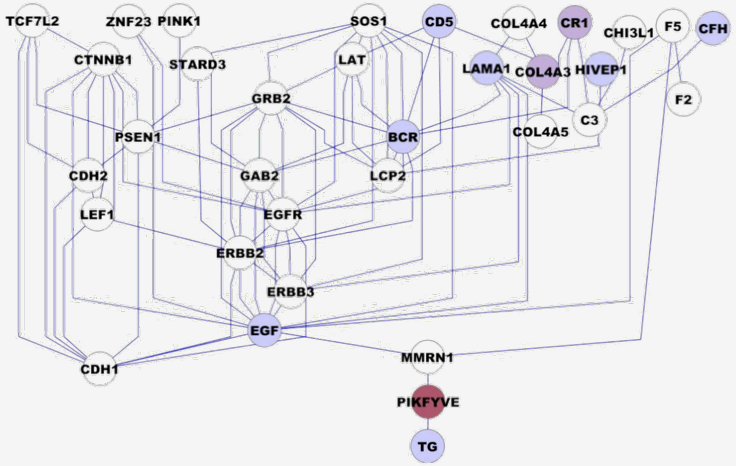
2.


TG
TRIM22
LEF1
THNSL2
PIKFYVE
DNAH17
MYO18B
CDH2
ZNF23
CEP120
TCF7L2
C15orf42
ARHGEF37
MMRN1
FAT2
MYEOV
PINK1
CTNNB1
CDH1
PSEN1
F5
F2
DDX59
CCNL2
EGF
NEFH
ERBB2
FAM20A
PMS2
EGFR
OVCH1
KIAA1731
STARD3
GRB2
PIK3C2G
GAB2
ERBB3
TRPA1
IGFN1
KCNK10
RNF215
SOS1
CD248
OR2V2
OR2B11
LCP2
BCR
LAMA1
ABCA10
LAT
TNN
HIVEP1
ARAP3
C7orf62
COL4A4
RPGRIP1
ACSBG1
ART1
COL4A3
CD5
C3
ACAN
SLC22A24
ABCA6
COL4A5
CR1
TTLL5
ANKRD11
CFH
TPPP2
DCAF6
LRMP
SMTN
DOPEY2
APOL3
IMMT
CHI3L1

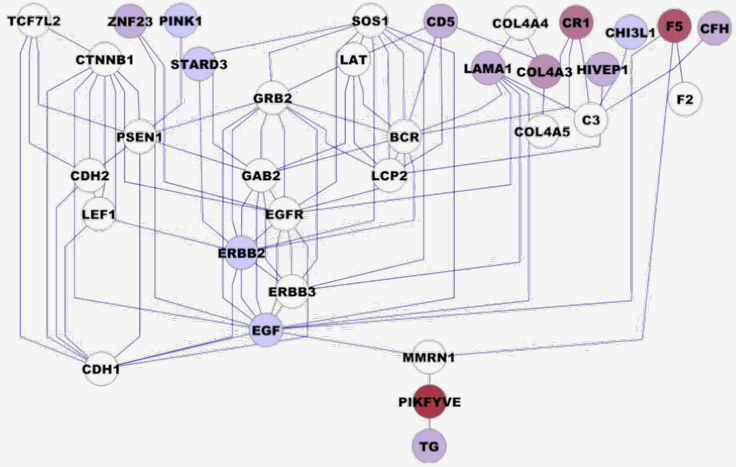
Homozygous SNVs Mother



Homozygous SNVs Father



Homozygous SNVs Children



```
GCTTCAGAACCCATCCATGTGAGGAAGTATAAAGGGCAGGTAGTAGCTGT
GCTTCGGAGCCCATGCATGTGAAGAAGTATCGAGGACAGACAGTGGCTGT
***** ** ***** ******* *******  *** ***  *** *****

GGATACATATTGCTGGCTTCACAAAGGAGCTATTGCTTGTGCTGAAAAAC
GGACACATACTGCTGGCTTCATAAAGGAGCTTTTTCATGTGCAGAGAAGC
*** ***** ********** ********* ** * ***** ** ** *

TAGCCAAAGGTGAACCTACTGATAGGTATGTAGGATTTTGTATGAAATTT
TTGCAAAAGGGGAACCTACAGATCAGTATGTCTCCTACTGTATGAAGTTT
* ** ***** ******** ***  ******    *  ******** ***

GTAAATATGTTACTATCTCATGGGATCAAGCCTATTCTCGTATTTGATGG
GTGGACATGCTGCTTTCTTTTGGTGTTAAACCTATCTTGGTGTTTGATGG
**  * *** * ** ***  ***  * ** *****  * ** ********

ATGTACTTTACCTTCTAAAAAGGAAGTAGAGAGATCTAGAAG
TCGTAACTTGCCCTCCAAACAGGAAGTGGAGAAGTCCCGGCG
  ***  ** ** ** *** ******* ****  **  *  *

GACAAGCCAATCTTCTTAAGGGAAAGCAACTTCTTCGTGAGG
GACAGGCCAATCTGCAGAAAGGCAAACAGCTGCTGCGGGAGG
**** ******** *  ** ** ** ** ** ** ** ****
```

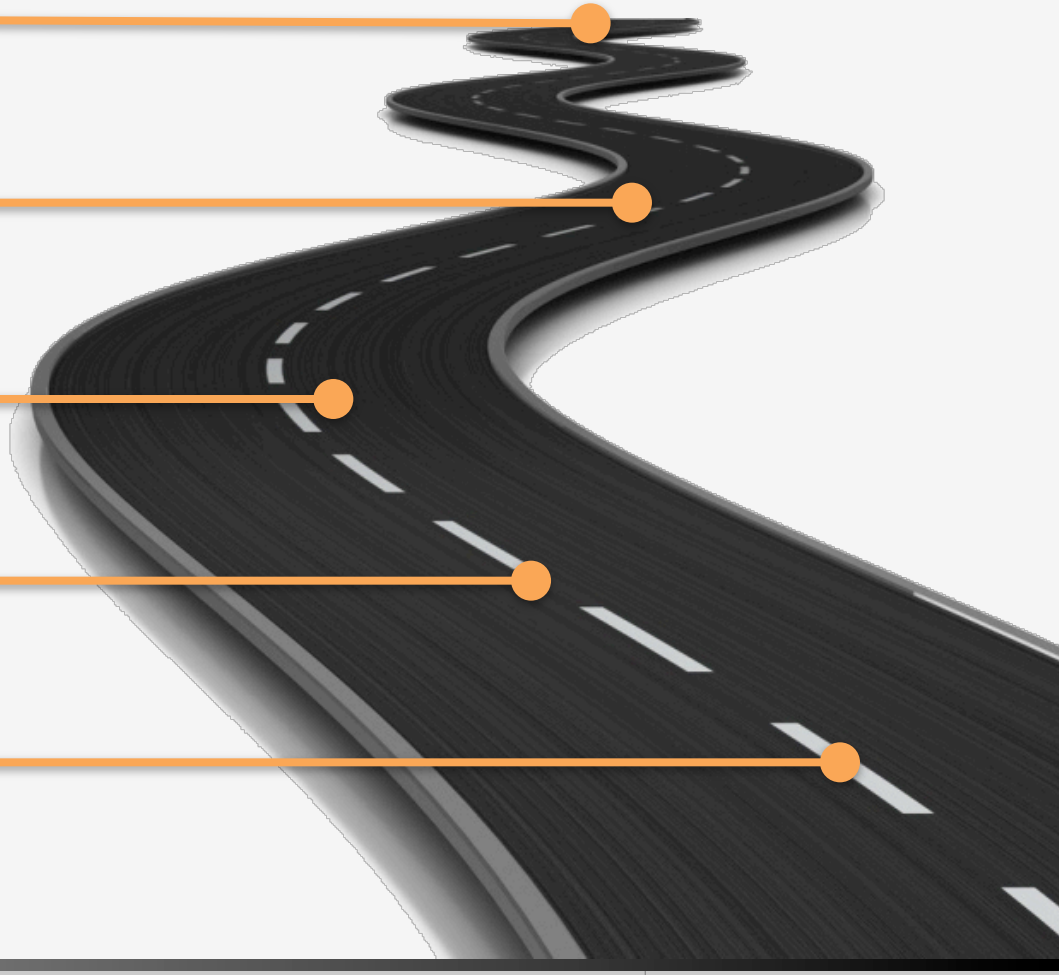
Introduction

Genome Sequencing

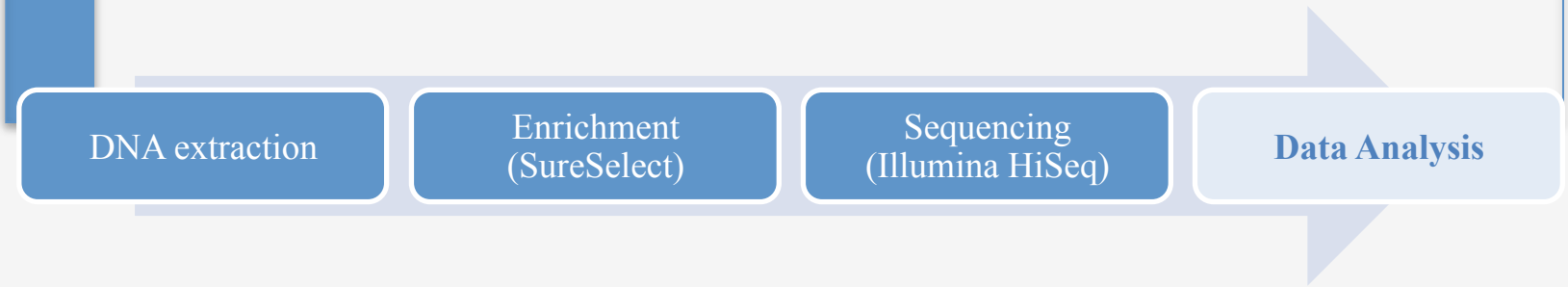
Exome Sequencing

Gene Panel Sequencing

Other applications



- **Gene Panel Screening**
- We screened approx. 700 patients (dilated cardiomyopathies) for almost 100 genes making up 500,000 bases.
- Sequencing was performed on Illumina HiSeq and Illumina MiSeq instruments, respectively.
- Challenge: Develop a fully automated solution for clinicians to handle and interpret NGS data



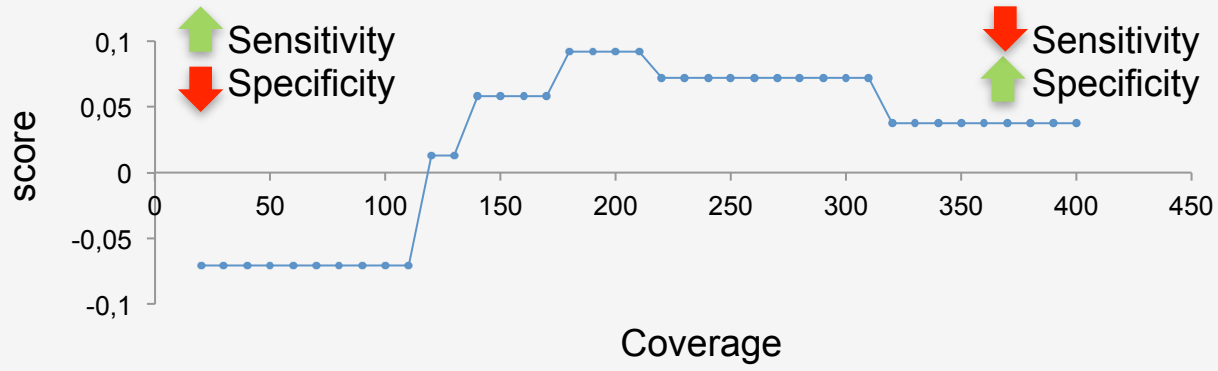
- Per patient roughly 2 billion bases are sequenced such that the 500,000 base region is covered on average maximal 4,000 fold
- About 99.5% of the total genomic region are covered at least with 50 reads to ensure diagnostic quality of genetic sequence

patients/coverage [%]	1x	5x	8x	10x	20x	50x	ADoC	% covered ≥20x	ADoC
patient1	99,99	99,95	99,92	99,91	99,82	99,58	3132,79	79.82	57.00
patient2	99,99	99,95	99,87	99,82	99,68	99,49	2150,17	75.98	43.63
patient3	99,99	99,99	99,99	99,98	99,90	99,61	2707,06	82.53	57.96
patient4	99,99	99,99	99,98	99,96	99,75	99,58	2279,10	82.23	59.95
patient5	99,99	99,99	99,98	99,98	99,79	99,51	2294,88	78.41	50.60
patient6	99,98	99,97	99,93	99,91	99,77	99,54	1770,30	79.10	51.82
patient7	99,94	99,79	99,71	99,66	99,49	99,10	1445,21		
patient8	99,99	99,94	99,93	99,92	99,78	99,56	2086,94		
patient9	99,99	99,95	99,89	99,87	99,66	99,41	1792,99		
patient10	99,99	99,95	99,91	99,86	99,72	99,44	2301,86		
mean	**99,99**	**99,95**	**99,91**	**99,89**	**99,74**	**99,48**		**79.68**	**53.49**

- One of the biggest challenges is still the SNP calling process. Analyzing wrong SNP calls we figured out 9 quality criteria that influence SNP calls significantly:
- Depth of coverage
- Allele balance
- Contiguous homopolymer run length
- Consistency with two segregating haplotypes
- 5 further related to the mapping quality including phred score

- Each of the parameters has been optimized separately on a training set using Matthews Correlation Coefficient (MCC). For filter *i* and parameter *j, MCC(i,j)** is calculated separately.

- The MCC offers a well suited measurement since it is well balanced and still useful if classes are of significantly different size

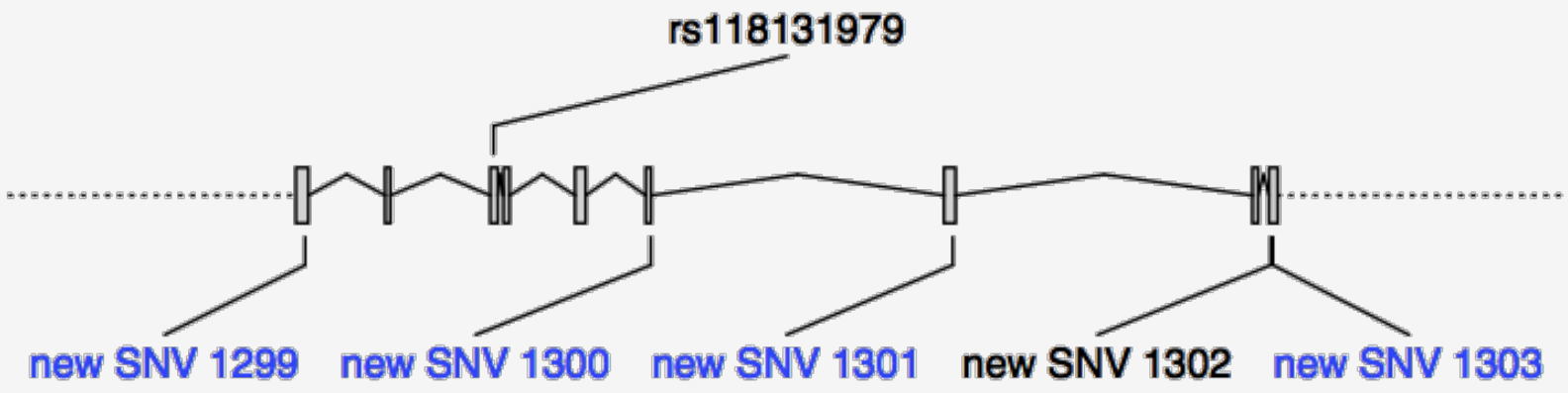


$$MCC_{(i,j)} = \frac{TP_{(i,j)} \times TN_{(i,j)} - FP_{(i,j)} \times FN_{(i,j)}}{\sqrt{(TP_{(i,j)} + FP_{(i,j)})(TP_{(i,j)} + FN_{(i,j)})(TN_{(i,j)} + FP_{(i,j)})(TN_{(i,j)} + FN_{(i,j)})}}$$

- Where $TP_{(i,j)}$ corresponds to the number of True Positive SNP calls for filter *i* and parameter *j*.
- $TN_{(i,j)}$, $FP_{(i,j)}$, $FN_{(i,j)}$ are defined analogously as True Negatives, False Positives and False Negatives for filter *i* and parameter *j*.

> Result per gene

SID547
SID146
SID388
SID173
SID163
SID916



SID278
SID293

SID914

SID908

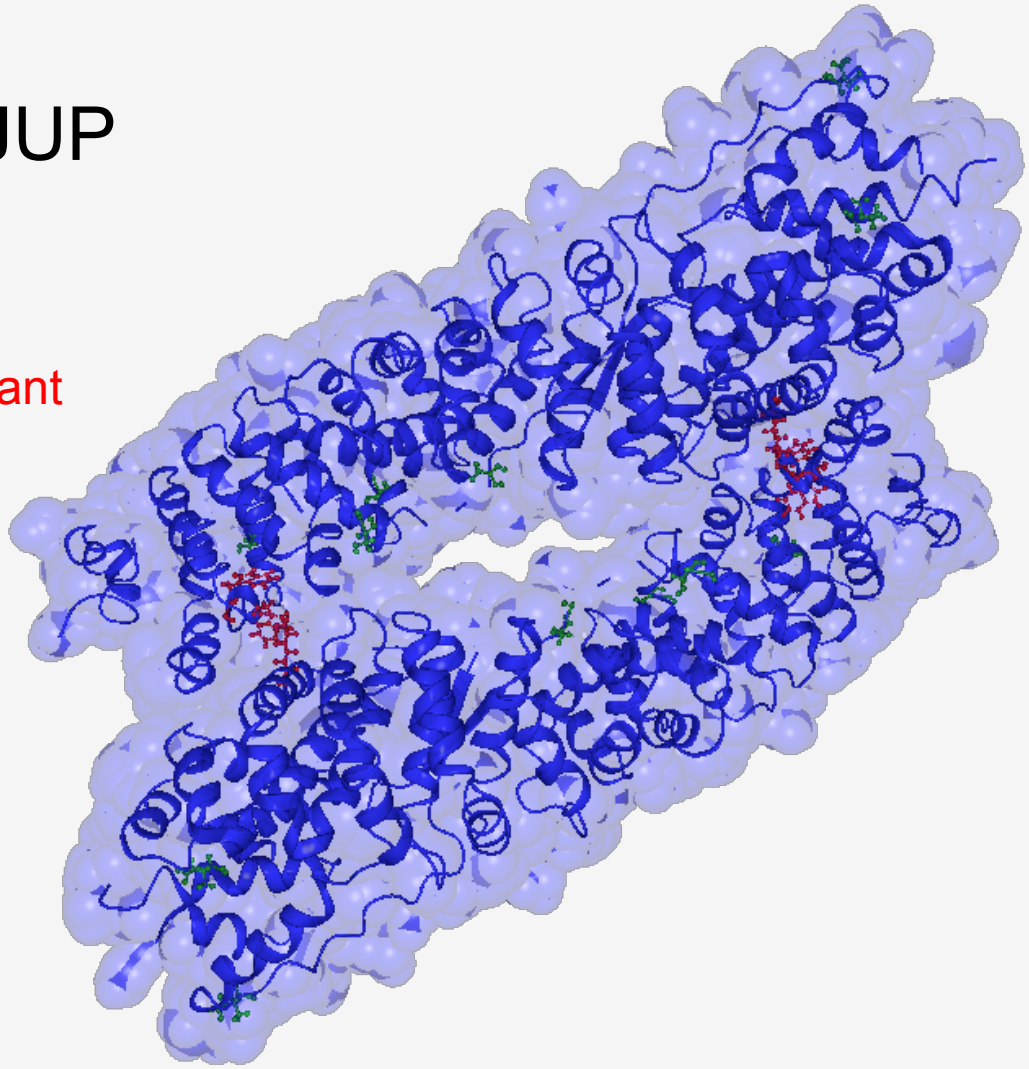
SID776

SID922

JUP

malignant

benign



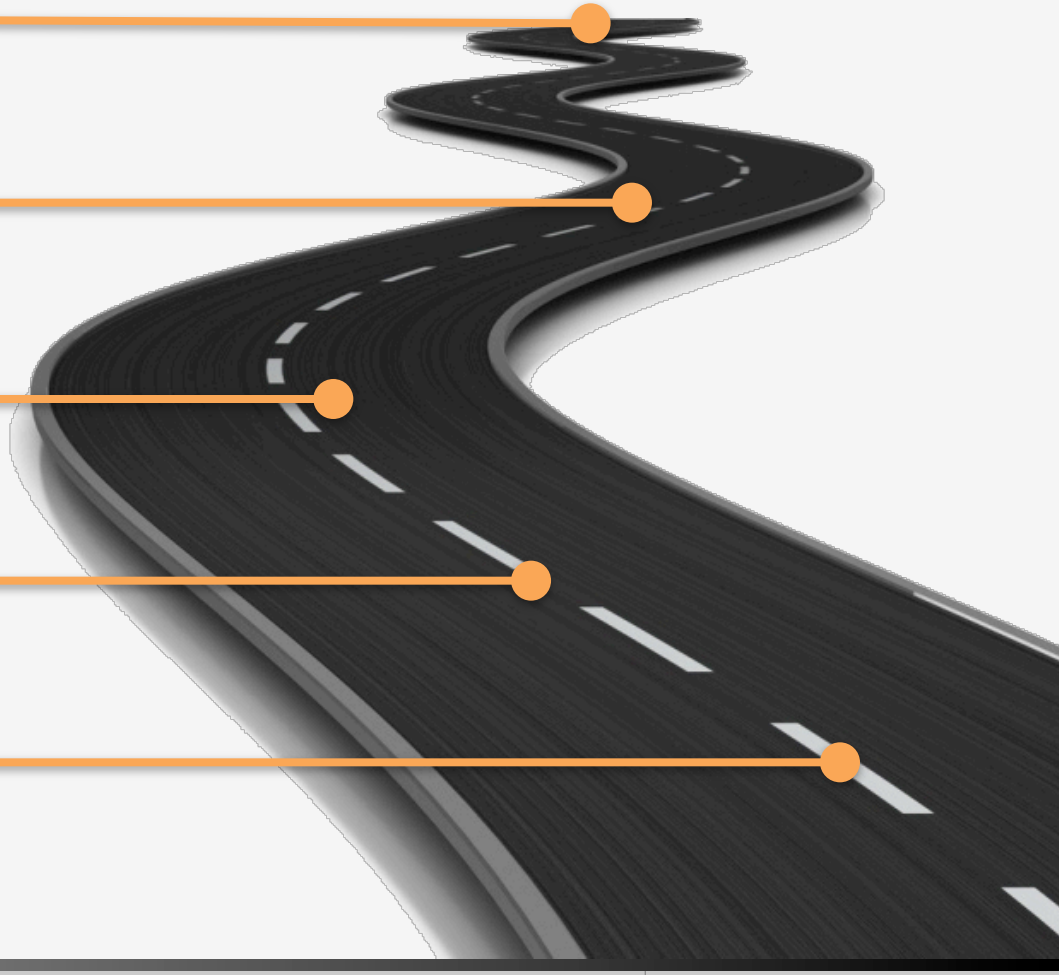
Introduction

Genome Sequencing

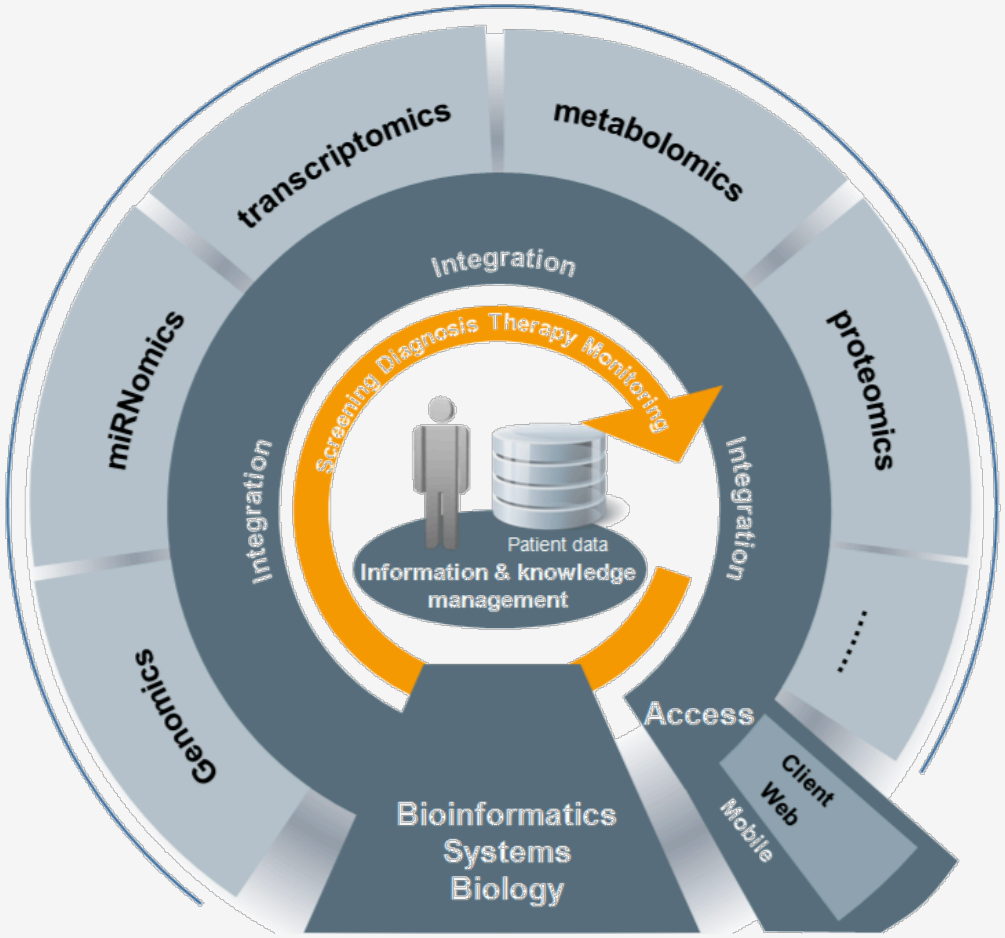
Exome Sequencing

Gene Panel Sequencing

Other applications



The biggest clinical value is in the integration of different patient data over a period of time



- **Complete characterization of patients with cardiac and neurological phenotypes**
- Whole Genome Sequencing of tissue
- Whole miRNome sequencing of blood, serum and tissue
- Transcriptome sequencing of tissue
- Methylation
- Targeted proteomics and metabolomics
- MRI Imaging

Application	Datset Size	# Datasets	Total Size
miRNA	5 GB	1,000	5,000
Gene Panels	10 GB	800	8,000
Transcriptomes	30 GB	100	3,000
Exomes	30 GB	500	15,000
Genomes	500 GB	120	60,000
Bacteria	5 GB	10,000	50,000
Together			1.4 PetaByte

Saarland University:
Prof. Eckart Meese
Prof. Hans-Peter Lenhof
Prof. Norbert Graf
Dr. Christina Backes
Dr. Petra Leidinger
Dr. Nicole Ludwig

Heidelberg University:
Prof. Hugo Katus
Dr. Benjamin Meder
Dr. Britta Vogel
Jan Haas
Karen Frese

DKFZ Heidelberg:
Dr. Jörg Hoheisel
Dr. Andrea Bauer

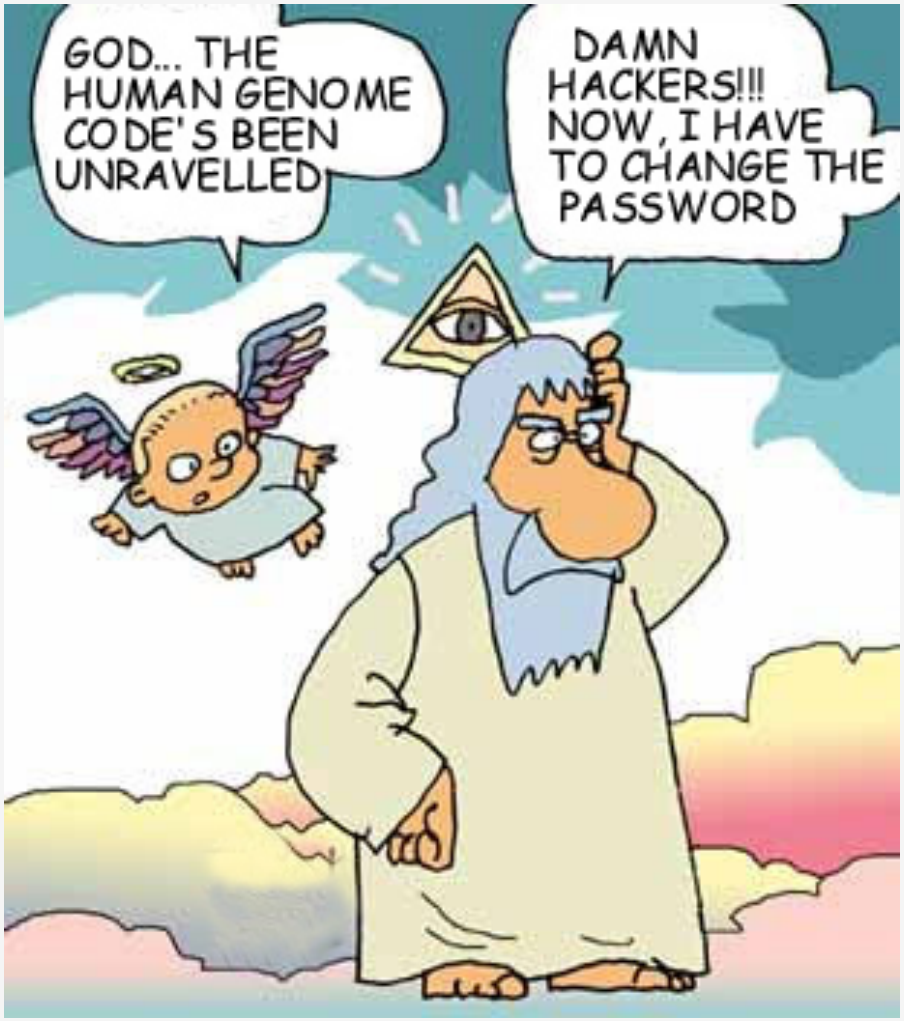
Kiel University
Prof. Schreiber
Prof. Andre Franke
Dr. Abdou ElSharawy
Dr. Michael Forster
Dr. Britt Peterson

Würzburg University
Prof. Dietel
Prof. Jörg Wischhusen
PD Dr. Sebastian Häusler

EURAC:
Prof. Albert Zink
Dr. Frank Maixner

Siemens:
Dr. Andreas Kappel
Dr. Jörn Mosner
Dr. Dorin Comaniciu
Dr. Emil Wirsz
Sarah Schlachter
Michal Skubacz
Cord Stähler
Jan Kirsten

CBC:
Dr. Markus Beier
Dr. Thomas Brefort
Jochen Kohlhaas


GOD... THE HUMAN GENOME CODE'S BEEN UNRAVELLED
DAMN HACKERS!!! NOW, I HAVE TO CHANGE THE PASSWORD