HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Analyze Genomes: An In-Memory Technology Use Case

In-Memory Computing for Life Sciences

Dr. Matthieu-P. Schapranow, Hasso Plattner Institute

Our Motivation
Personalized Medicine

- **Motivation:** Can we analyze the entire data of a patient, incl. Electronic Health Records (EHR) and genome data, during a doctor's visit?

- Genome data analysis may add up to weeks, i.e. biopsy, biological preparation, sequencing, alignment, variant calling, full analysis, and evaluation
- Issue: Complex and time-consuming data processing tasks

- In-memory technology accelerates genome data processing
  - □ Highly parallel alignment / variant calling
  - □ Real-time analysis of individual patient or cohort data
  - □ Combined search in structured / unstructured data

Our Challenge
Distributed Big Data Sources

Human genome/biological data
600GB per full genome
15PB+ in databases of leading institutes

Human proteome
160M data points (2.4GB) per sample
>3TB raw proteome data in ProteomicsDB

Hospital information systems
Often more than 50GB

PubMed database
>23M articles

Cancer patient records
>160k records at NCT

Medical sensor data
Scan of a single organ in 1s
creates 10GB of raw data

Prescription data
1.5B records from 10,000 doctors
and 10M Patients (100 GB)

Clinical trials
Currently more than 30k
recruiting on ClinicalTrials.gov

Our Approach
In-Memory Technology

Combined column and row store

Insert only for time travel

Active/passive data store

Dynamic multi-threading

No aggregate tables

On-the-fly extensibility

Map/Reduce

Real-time replication

Minimal projections

Bulk load of data

Data partitioning

Analytics on historical data

Single and multi-tenancy

Working on integers

Group key

Object-relational mapping

Any attribute as index

Multi-core/ parallelization

Lightweight compression

SQL interface on columns and rows

Reduction of software layers

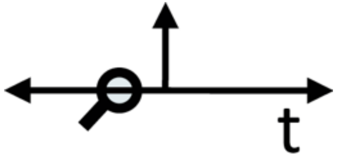
Text retrieval and extraction engine

No disk

Meaningful In-Memory Database Concepts
Text Mining

- Full text indexing for any text attributes
- User-defined **dictionaries** to define entities
  http://scn.sap.com/community/developer-center/hana/blog/2013/12/27/hana-text-analysis-with-custom-dictionaries
- Custom Grouper User Language (**CGUL**) rules to create token-based regular expressions with linguistic attributes
  http://wiki.scn.sap.com/wiki/display/EIM/CGUL+Tips+and+Tricks+for+Entity+Extraction

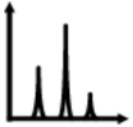
Meaningful In-Memory Database Concepts
Time Travel



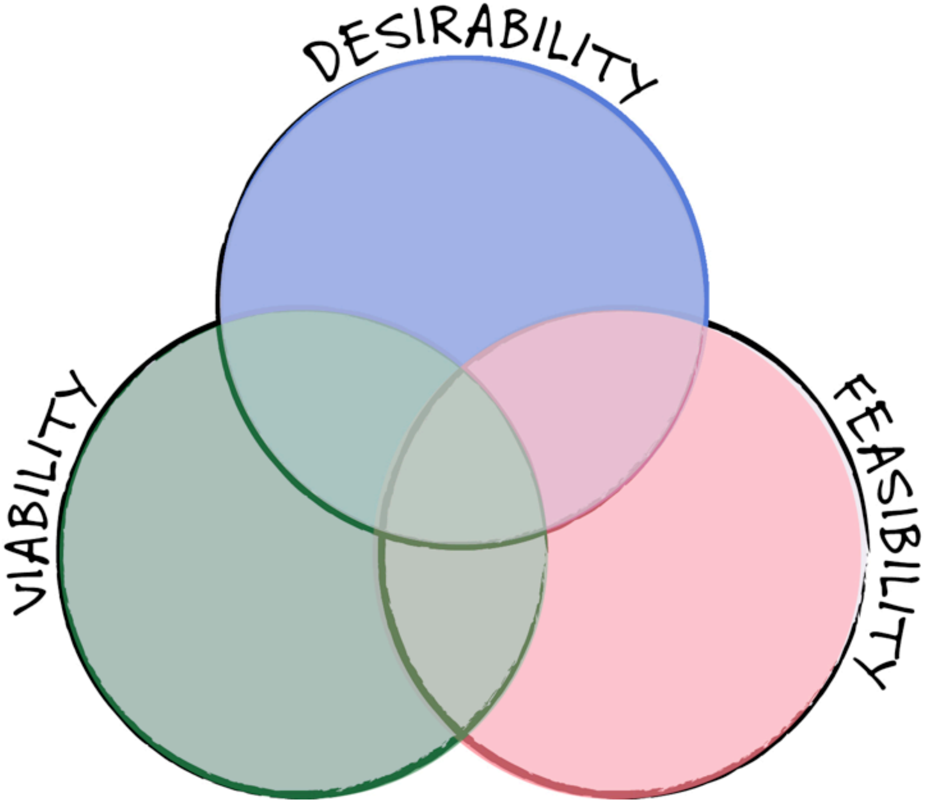
- Process time series data by retrieving the complete database state at any period of history
- **History Database Tables**
- SELECT TEMPERATURE FROM ”PATIENTS”.”ICS_TEMP” WHERE PATIENT NAME = ‘Matthieu Schapranow’ AS OF COMMIT ID 209811
- http://scn.sap.com/community/developer-center/hana/blog/2013/02/12/when-i-travelled-through-time-using-sap-hana

Meaningful In-Memory Database Concepts
Predictive Analysis Library (PAL)

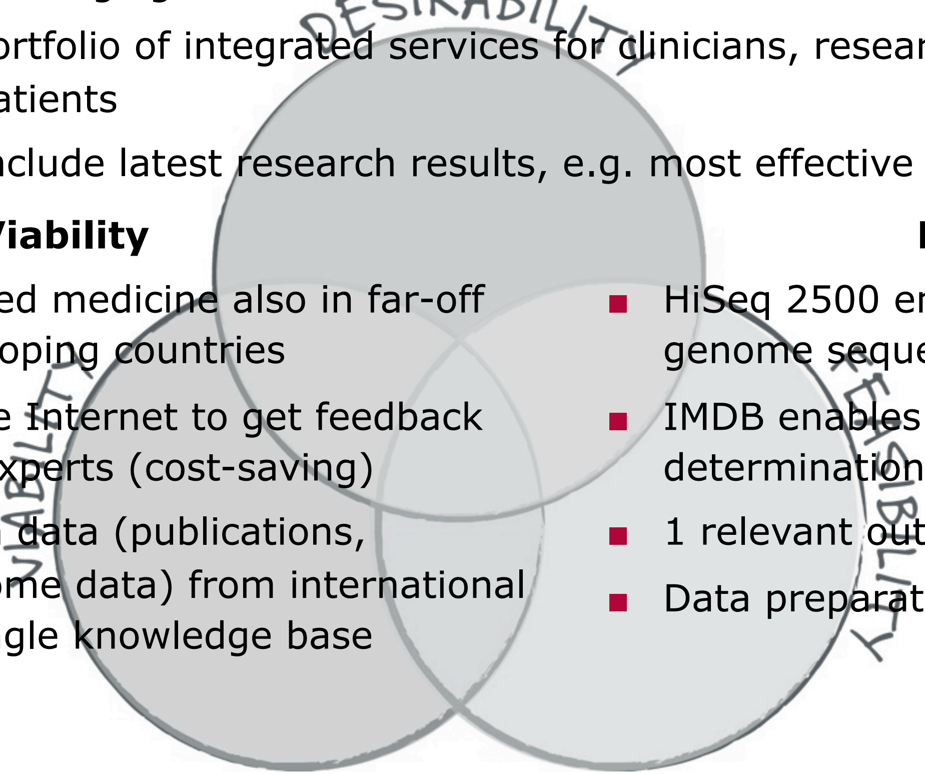
- Provides specific **analysis functions** tightly integrated within the database, e.g. k-means or hierarchical clustering
- http://help.sap.com/hana/SAP_HANA_Predictive_Analysis_Library_PAL_en.pdf



Our Vision
Personalized Medicine



Our Vision
Personalized Medicine



Desirability

- Leveraging directed customer services
- Portfolio of integrated services for clinicians, researchers, and patients
- Include latest research results, e.g. most effective therapies

Viability

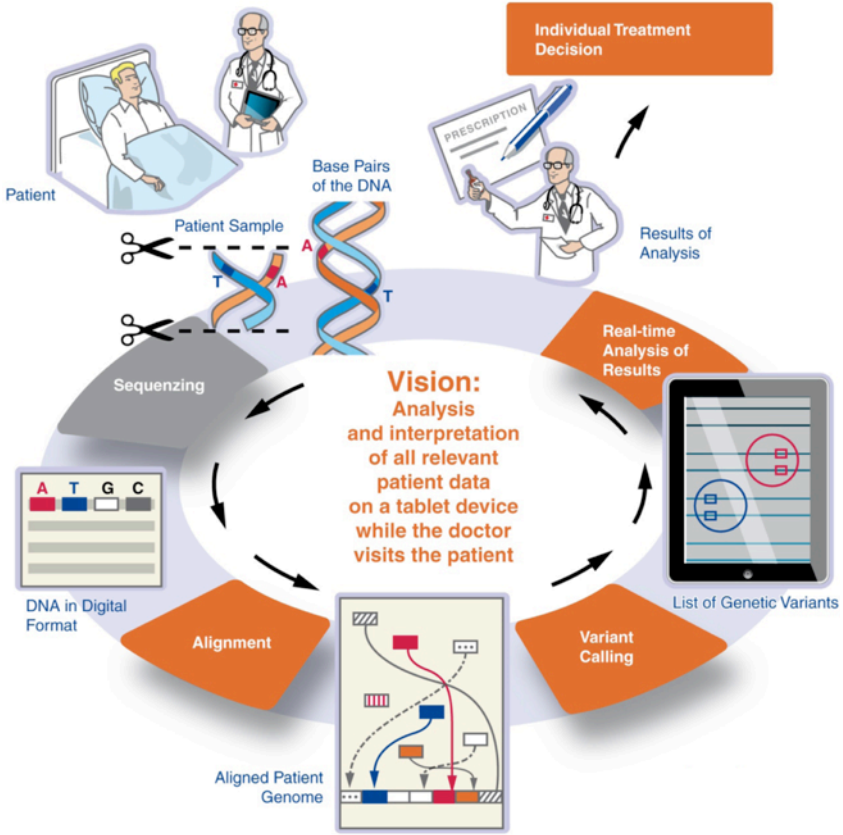
- Enable personalized medicine also in far-off regions and developing countries
- Share data via the Internet to get feedback from word-wide experts (cost-saving)
- Combine research data (publications, annotations, genome data) from international databases in a single knowledge base

Feasibility

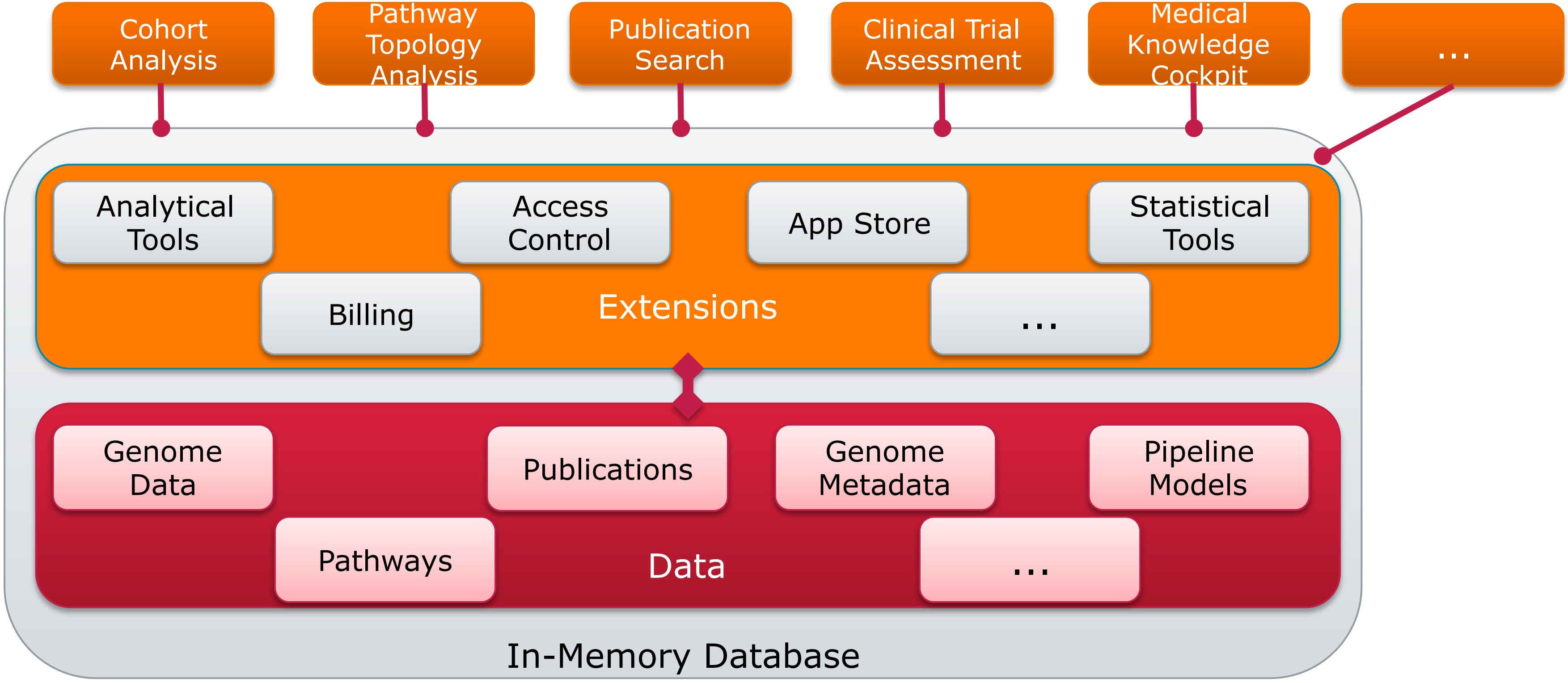
- HiSeq 2500 enables high-coverage whole genome sequencing in ≈1d
- IMDB enables allele frequency determination of 12B records within <1s
- 1 relevant out of 80M annotations <1s
- Data preparation as a service reduces TCO

High-Performance In-Memory Genome Project
Integration of Genomic Data

- Preprocessing of DNA (alignment, variant calling) can be modeled and is executed as integrated process
- Results are directly stored in in-memory databases, e.g. for
  - □ Statistical analyses, and
  - □ Links to latest research knowledge
- Real-time analysis of genome data enables completely new way of research and therapies, e.g. instant comparison with patient cohorts



High-Performance In-Memory Genome Project Architectural Overview



High-Performance In-Memory Genome Project
Selected Research Topics

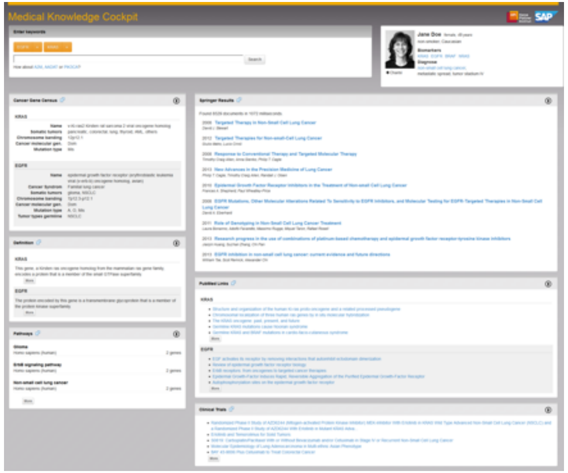
Improving Analyses:

- Information combination, e.g. Medical Knowledge Cockpit, Oncolyzer
- Genome Browser enables deep dive into the genome
- Cohort Analysis, e.g. clustering of patient cohorts
- Combined Search, e.g. in clinical trials and side-effect databases
- Pathways Topology Analysis, e.g. to identify cause/effect

Improving Data Preparations:

- Graphical modeling of Genome Data Processing (GDP) pipelines
- Scheduling and execution of multiple GPD pipelines in parallel
- App store for medical knowledge (bring algorithms to data)
- Exchange of sensitive data, e.g. history-based access control
- Billing processes for intellectual property and services

High-Performance In-Memory Genome Project
Medical Knowledge Cockpit



Medical Knowledge Cockpit

Enter keywords

Search

Jane Doe female

non-smoker, Caucasian

Biomarkers

KRAS EGFR BRAF NRAS

Diagnose

non-small cell lung cancer,

metastatic spread, tumor stadium IV

Cancer Gene Census

Springer Results

2008 Targeted Therapy in Non-Small Cell Lung Cancer

2012 Targeted Therapies for Non-small-Cell Lung Cancer

2008 Response to Conventional Therapy and Targeted Molecular Therapy

2013 New Advances in the Precision Medicine of Lung Cancer

2010 Epidermal Growth Factor Receptor Inhibitors in the Treatment of Non-small Cell Lung Cancer

2008 EGFR Mutations, Other Molecular Alterations Related To Sensitivity to EGFR Inhibitors, and Molecular Testing for EGFR-Targeted Therapies in Non-Small Cell Lung Cancer

2011 Role of Genotyping in Non-Small Cell Lung Cancer Treatment

2013 Research progress in the use of combinations of platinum-based chemotherapy and epidermal growth factor receptor-tyrosine kinase inhibitors

2013 EGFR inhibition in non-small cell lung cancer: current evidence and future directions

Definition

Pathways

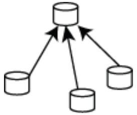
PubMed Links

KRAS

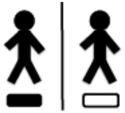
- Structure and organization of the human Ki-ras proto-oncogene and a related processed pseudogene
- Chromosomal localization of three human ras genes by in situ molecular hybridization
- The KRAS oncogene: past, present, and future

EGFR

Clinical Trials



Unified access to structured and un-structured data sources

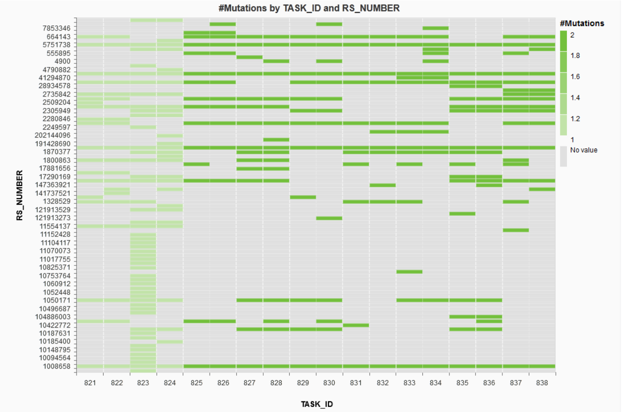


Automatic clinical trial matching build on text analysis features

- Search for affected genes in distributed and heterogeneous data sources
- Immediate exploration of relevant information, such as
  - □ Gene descriptions,
  - □ Molecular impact and related pathways,
  - □ Scientific publications, and
  - □ Suitable clinical trials.

- No manual searching for hours or days: In-memory technology translates searching into interactive finding!

High-Performance In-Memory Genome Project
Drug Response Testing

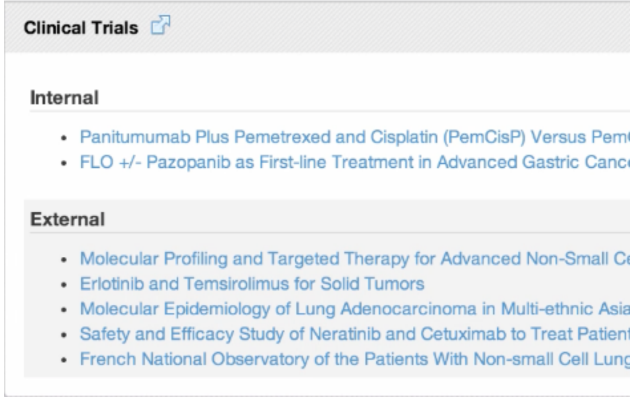


Interactive analysis of correlations between drugs and genetic variants

- Drug response depends on individual genetic variants of tumors
- Challenge: Identification of relevant genetic variants and their impact on drug response is a ongoing research activity, e.g. Xenograft models
- Exploration of experiment results is time-consuming and Excel-driven
- In-memory technology enables interactive exploration of experiment data to leverage new scientific insights

High-Performance In-Memory Genome Project

Search in Structured and Unstructured Medical Data



Clinical Trials

Internal

- Panitumumab Plus Pemetrexed and Cisplatin (PemCisP) Versus Pem
- FLO +/- Pazopanib as First-line Treatment in Advanced Gastric Canc

External

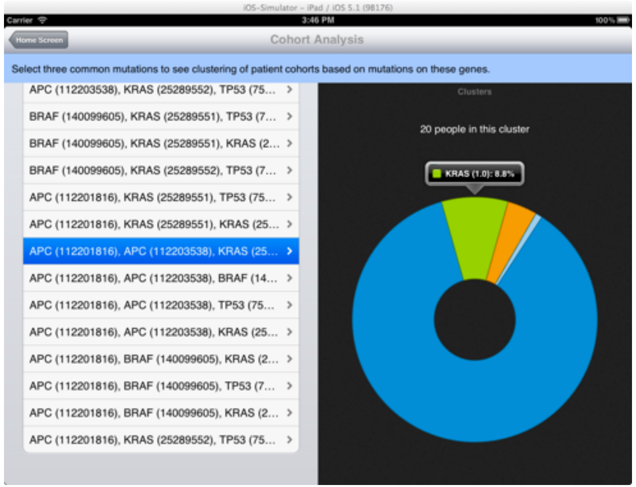
- Molecular Profiling and Targeted Therapy for Advanced Non-Small Ce
- Erlotinib and Temsirolimus for Solid Tumors
- Molecular Epidemiology of Lung Adenocarcinoma in Multi-ethnic Asia
- Safety and Efficacy Study of Neratinib and Cetuximab to Treat Patient
- French National Observatory of the Patients With Non-small Cell Lung

Unified access to structured and unstructured data sources

Clinical trial matching using text analysis features

- Extended text analysis feature by medical terminology
  - □ Genes (122,975 + 186,771 synonyms)
  - □ Medical terms and categories (98,886 diseases + 48,561 synonyms, 47 categories)
  - □ Pharmaceutical ingredients (7,099 + 5,561 synonyms)
- Indexed clinicaltrials.gov database (145k trials/ 30,138 recruiting)
- Extracted, e.g., 320k genes, 161k ingredients, 30k periods
- Select all studies based on multiple filters in less than 500ms

High-Performance In-Memory Genome Project
Analysis of Patient Cohorts



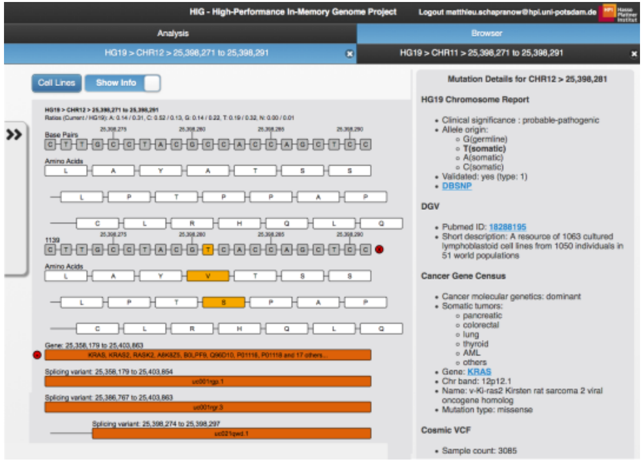
Select three common mutations to see clustering of patient cohorts based on mutations on these genes.

- APC (112203538), KRAS (25289552), TP53 (75...
- BRAF (140099605), KRAS (25289551), TP53 (7...
- BRAF (140099605), KRAS (25289551), KRAS (2...
- BRAF (140099605), KRAS (25289552), TP53 (7...
- APC (112201816), KRAS (25289551), TP53 (75...
- APC (112201816), KRAS (25289551), KRAS (25...
- APC (112201816), APC (112203538), KRAS (25...
- APC (112201816), APC (112203538), BRAF (14...
- APC (112201816), APC (112203538), TP53 (75...
- APC (112201816), APC (112203538), KRAS (25...
- APC (112201816), BRAF (140099605), KRAS (2...
- APC (112201816), BRAF (140099605), TP53 (7...
- APC (112201816), BRAF (140099605), KRAS (2...
- APC (112201816), KRAS (25289552), TP53 (75...

Fast clustering directly performed within the in-memory database

- In a patient cohort, a subset does not respond to therapy – why?
- Clustering using various statistical algorithms, such as k-means or hierarchical clustering
- Calculation of all locus combinations in which at least 5% of all TCGA participants have mutations: 200ms for top 20 combinations
- Individual clusters are calculated in parallel directly within the database
- K-means algorithm: 50ms (PAL) vs. 500ms (R)

High-Performance In-Memory Genome Project
Genome Browser


HIG - High-Performance In-Memory Genome Project
Logout matthieu.schapranow@hpi.uni-potsdam.de
Analysis
Browser
HG19 > CHR12 > 25,398,271 to 25,398,291
HG19 > CHR11 > 25,398,271 to 25,398,291
Cell Lines
Show Info
Base Pairs
Amino Acids
25,398,275
25,398,280
25,398,285
25,398,290

HG19 > CHR12 > 25,398,271 to 25,398,291
Ratios (Current / HG19): A: 0.14 / 0.31, C: 0.52 / 0.13, G: 0.14 / 0.22, T: 0.19 / 0.32, N: 0.00 / 0.01

Gene: 25,358,179 to 25,403,863
KRAS, KRAS2, RASK2, A6K8Z5, B0LPF9, Q96D10, P01116, P01118 and 17 others...

Splicing variant: 25,358,179 to 25,403,854
uc001rgp.1

Splicing variant: 25,386,767 to 25,403,863
uc001rgr.3

Splicing variant: 25,398,274 to 25,398,297
uc021qwd.1

Mutation Details for CHR12 > 25,398,281

HG19 Chromosome Report

- Clinical significance : probable-pathogenic
- Allele origin:
  - G(germline)
  - **T(somatic)**
  - A(somatic)
  - C(somatic)
- Validated: yes (type: 1)
- DBSNP

DGV

- Pubmed ID: 18288195
- Short description: A resource of 1063 cultured lymphoblastoid cell lines from 1050 individuals in 51 world populations

Cancer Gene Census

- Cancer molecular genetics: dominant
- Somatic tumors:
  - pancreatic
  - colorectal
  - lung
  - thyroid
  - AML
  - others
- Gene: KRAS
- Chr band: 12p12.1
- Name: v-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog
- Mutation type: missense

Cosmic VCF

- Sample count: 3085

Unified access to multiple formerly disjoint data sources

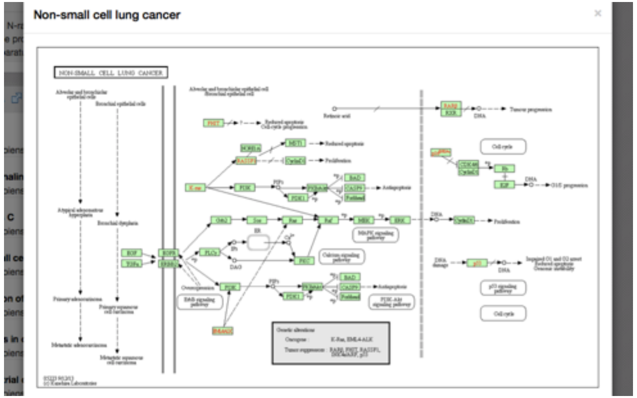
Matching of genetic variants and relevant annotations

- Genome Browser: Comparison of multiple genomes with reference
- Combined knowledge base: latest relevant annotations and literature, e.g. NCBI, dbSNP, and UCSC

- Detailed exploration of genome locations and existing associations
- Ranked variants, e.g. accordingly to known diseases

- Links to more detailed sources enable fast identification of relevant information while eliminating long-lasting searches.

High-Performance In-Memory Genome Project
Pathway Analysis



- Search in pathways is limited to “is a certain element contained” today
- Integrated >1,5k pathways from international sources, e.g. KEGG, HumanCyc, and WikiPathways, into HANA
- Implemented graph-based topology exploration and ranking based on patient specifics
- Enables interactive identification of possible dysfunctions affecting the course of a therapy before its start

Unified access to multiple formerly disjoint data sources

Pathway analysis of genetic variants with graph engine

What to take home?
Test it: http://we.AnalyzeGenomes.com

For researchers

- Enable real-time analysis of medical data
- Automatic assessment of data, e.g. scan of pathways to identify cellular impact of mutations
- Combined free-text search in publications, diagnosis, and EMR data, i.e. structured and unstructured data

For clinicians

- Preventive diagnostics to identify risk patients early
- Indicate pharmacokinetic correlations
- Scan for similar patient cases, e.g. to evaluate therapy success

For patients

- Identify relevant clinical trials and medical experts
- Start most appropriate therapy as early as possible

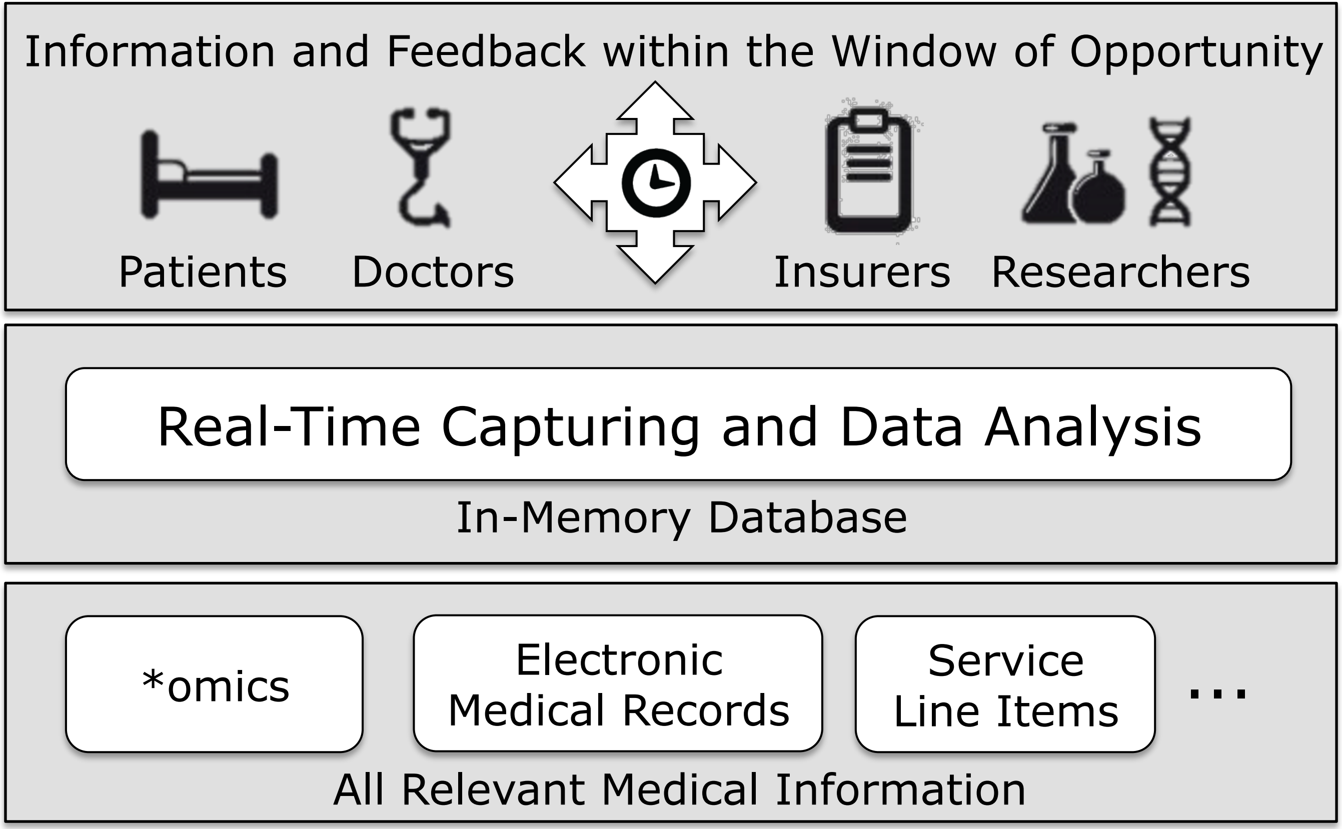
Keep in contact with us!

Dr. Matthieu-P. Schapranow
schapranow@hpi.uni-potsdam.de
http://we.analyzegenomes.com/

Hasso Plattner Institute
Enterprise Platform & Integration Concepts
Dr. Matthieu-P. Schapranow
August-Bebel-Str. 88
14482 Potsdam, Germany

BACKUP

High-Performance In-Memory Genome Project
Architectural Overview



Information and Feedback within the Window of Opportunity

Patients

Doctors

Insurers

Researchers

Real-Time Capturing and Data Analysis

In-Memory Database

*omics

Electronic Medical Records

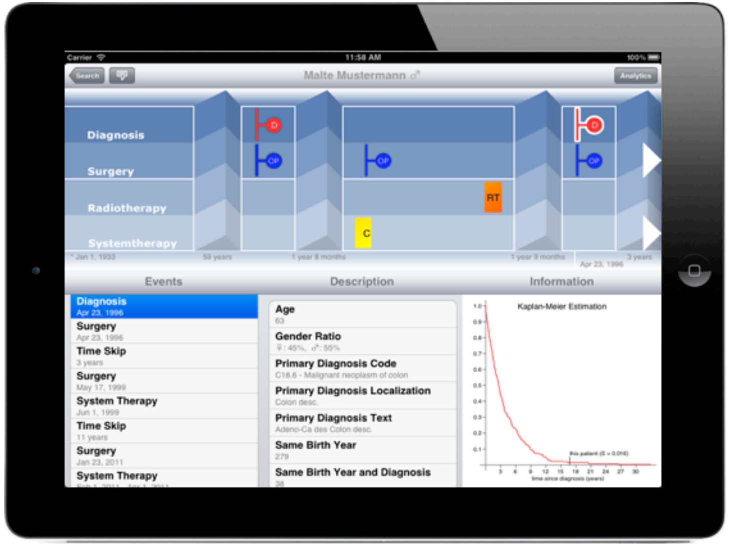
Service Line Items

...

All Relevant Medical Information

High-Performance In-Memory Genome Project
HANA Oncolyzer

- Research initiative for exchanging relevant tumor data to improve personalized treatment
- Honored 2012 by the Innovation Award of the German Capitol Region
- In-memory technology as key-enabler for real-time analysis of tumor data in seconds instead of hours
- Information available at your fingertips: In-memory technology on mobile devices, e.g. iPad
- Interdisciplinary cooperation between
  - □ Medical doctors,
  - □ Researchers, and
  - □ Software engineers.

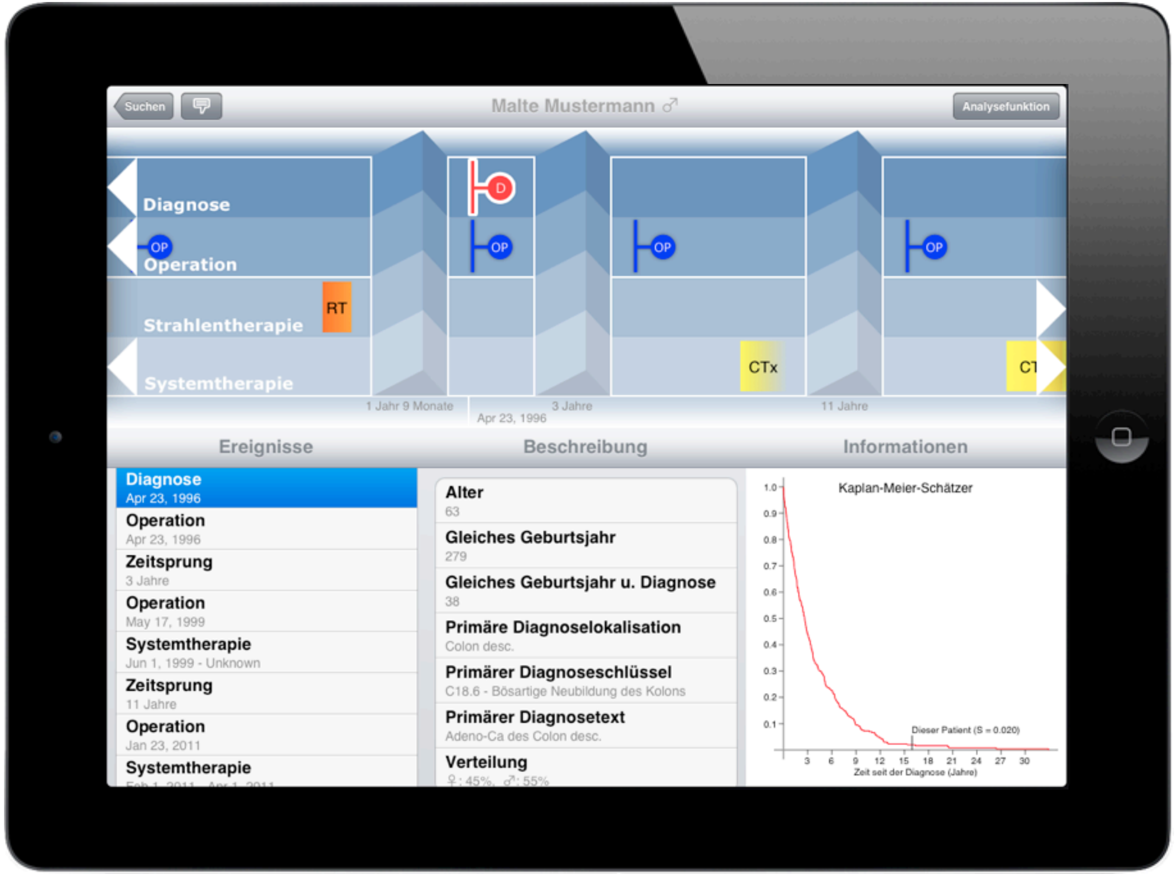


Diagnosis
Apr 23, 1996
Surgery
Apr 23, 1996
Time Skip
3 years
Surgery
May 17, 1999
System Therapy
Jun 1, 1999
Time Skip
11 years
Surgery
Jan 23, 2011
System Therapy
Feb 1, 2011 - Apr 1, 2011

Age
63
Gender Ratio
♀: 45%, ♂: 55%
Primary Diagnosis Code
C18.6 - Malignant neoplasm of colon
Primary Diagnosis Localization
Colon desc.
Primary Diagnosis Text
Adeno-Ca des Colon desc.
Same Birth Year
279
Same Birth Year and Diagnosis
38

High-Performance In-Memory Genome Project HANA Oncolyzer Patient Details Screen

- Combines patient’s time series data of specific patient and analysis results of patient cohort
- Real-time analysis across hospital-wide data whenever details screen is accessed
- http://epic.hpi.uni-potsdam.de/Home/HanaOncolyzer



Diagnose
Apr 23, 1996

Operation
Apr 23, 1996

Zeitsprung
3 Jahre

Operation
May 17, 1999

Systemtherapie
Jun 1, 1999 - Unknown

Zeitsprung
11 Jahre

Operation
Jan 23, 2011

Systemtherapie
Feb 1, 2011 - Apr 1, 2011

Alter
63

Gleiches Geburtsjahr
279

Gleiches Geburtsjahr u. Diagnose
38

Primäre Diagnoselokalisation
Colon desc.

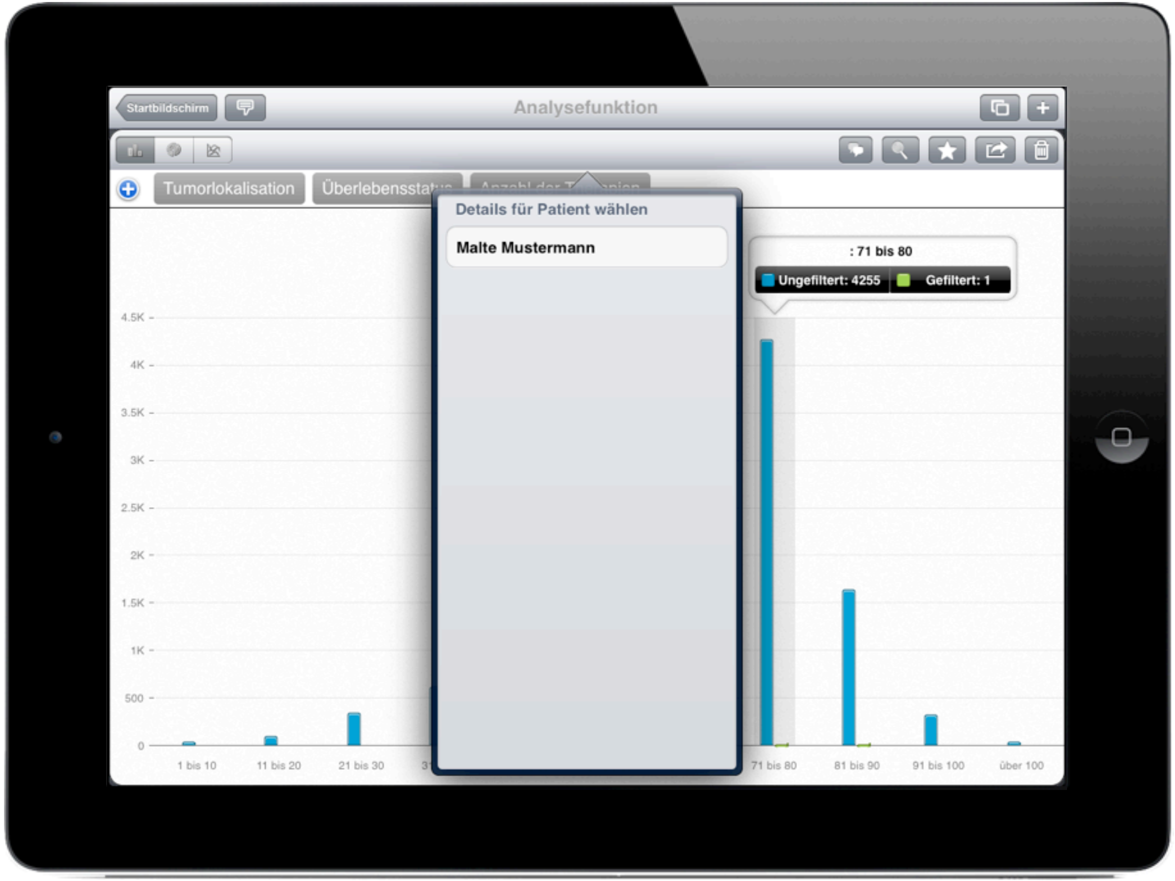
Primärer Diagnoseschlüssel
C18.6 - Bösartige Neubildung des Kolons

Primärer Diagnosetext
Adeno-Ca des Colon desc.

Verteilung
♀: 45%, ♂: 55%

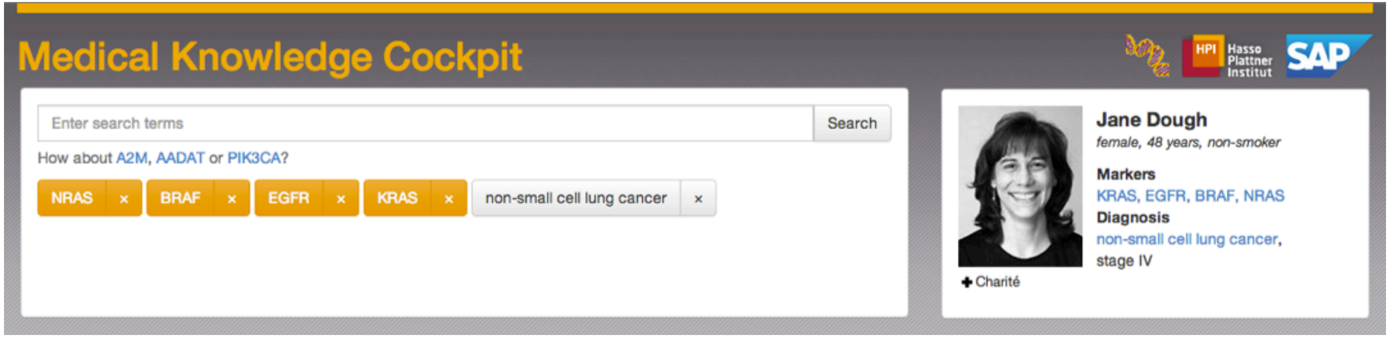
High-Performance In-Memory Genome Project
HANA Oncolyzer
Patient Analysis Screen

- Allows to real-time analysis on complete patient cohort
- Flexible filters and various chart types allow graphical exploration of data on mobile devices



Medical Knowledge Cockpit

Seamless Integration of Patient Specifics


Medical Knowledge Cockpit
HPI Hasso Plattner Institut
SAP
Enter search terms
Search
How about A2M, AADAT or PIK3CA?
NRAS ×
BRAF ×
EGFR ×
KRAS ×
non-small cell lung cancer ×
Charité

Jane Dough

female, 48 years, non-smoker

Markers

KRAS, EGFR, BRAF, NRAS

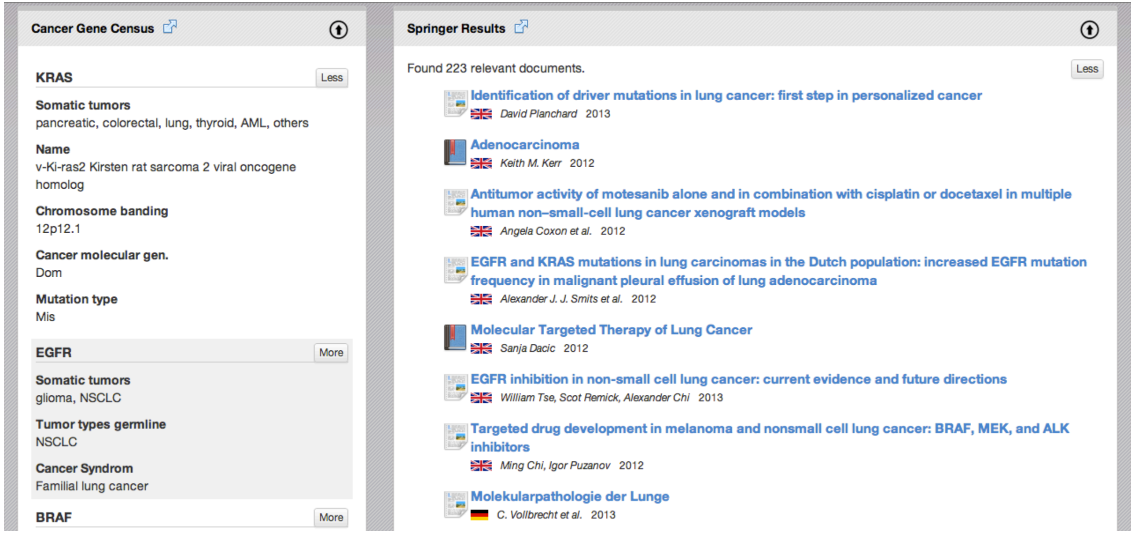
Diagnosis

non-small cell lung cancer, stage IV

- Google-like user interface for searching data
- Seamless integration of individual EMR data
- Search various sources for biomarkers, literature, and diseases

Medical Knowledge Cockpit

Publications



Cancer Gene Census

KRAS

Somatic tumors
pancreatic, colorectal, lung, thyroid, AML, others

Name
v-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog

Chromosome banding
12p12.1

Cancer molecular gen.
Dom

Mutation type
Mis

EGFR

Somatic tumors
glioma, NSCLC

Tumor types germline
NSCLC

Cancer Syndrom
Familial lung cancer

BRAF

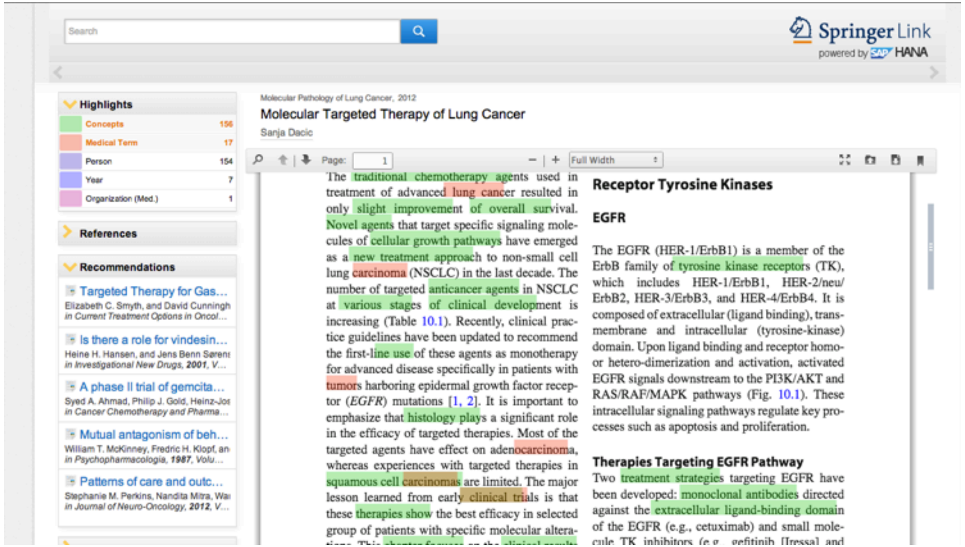
Springer Results

Found 223 relevant documents.

- Identification of driver mutations in lung cancer: first step in personalized cancer
  David Planchard 2013
- Adenocarcinoma
  Keith M. Kerr 2012
- Antitumor activity of motesanib alone and in combination with cisplatin or docetaxel in multiple human non–small-cell lung cancer xenograft models
  Angela Coxon et al. 2012
- EGFR and KRAS mutations in lung carcinomas in the Dutch population: increased EGFR mutation frequency in malignant pleural effusion of lung adenocarcinoma
  Alexander J. J. Smits et al. 2012
- Molecular Targeted Therapy of Lung Cancer
  Sanja Dacic 2012
- EGFR inhibition in non-small cell lung cancer: current evidence and future directions
  William Tse, Scot Remick, Alexander Chi 2013
- Targeted drug development in melanoma and nonsmall cell lung cancer: BRAF, MEK, and ALK inhibitors
  Ming Chi, Igor Puzanov 2012
- Molekularpathologie der Lunge
  C. Vollbrecht et al. 2013

- In-place preview of relevant data, such as publications and publication meta data
- Incorporating individual filter settings, e.g. additional search terms

Medical Knowledge Cockpit
Publications



Search

Highlights

🟩	Concepts	156
🟥	Medical Term	17
🟪	Person	154
🟦	Year	7
🟪	Organization (Med.)	1

References

Recommendations

- Targeted Therapy for Gas…
  Elizabeth C. Smyth, and David Cunningh
  in Current Treatment Options in Oncol…
- Is there a role for vindesin…
  Heine H. Hansen, and Jens Benn Sørens
  in Investigational New Drugs, ***2001****, V…*
- A phase II trial of gemcita…
  Syed A. Ahmad, Philip J. Gold, Heinz-Jos
  in Cancer Chemotherapy and Pharma…
- Mutual antagonism of beh…
  William T. McKinney, Fredric H. Klopf, an
  in Psychopharmacologia, ***1987****, Volu…*
- Patterns of care and outc…
  Stephanie M. Perkins, Nandita Mitra, Wa
  in Journal of Neuro-Oncology, ***2012****, V…*

Molecular Pathology of Lung Cancer, 2012

Molecular Targeted Therapy of Lung Cancer

Sanja Dacic

Page: 1

Full Width

The traditional chemotherapy agents used in treatment of advanced lung cancer resulted in only slight improvement of overall survival. Novel agents that target specific signaling molecules of cellular growth pathways have emerged as a new treatment approach to non-small cell lung carcinoma (NSCLC) in the last decade. The number of targeted anticancer agents in NSCLC at various stages of clinical development is increasing (Table 10.1). Recently, clinical practice guidelines have been updated to recommend the first-line use of these agents as monotherapy for advanced disease specifically in patients with tumors harboring epidermal growth factor receptor (*EGFR*) mutations [1, 2]. It is important to emphasize that histology plays a significant role in the efficacy of targeted therapies. Most of the targeted agents have effect on adenocarcinoma, whereas experiences with targeted therapies in squamous cell carcinomas are limited. The major lesson learned from early clinical trials is that these therapies show the best efficacy in selected group of patients with specific molecular alterations. This chapter focuses on the clinical results

Receptor Tyrosine Kinases

EGFR

The EGFR (HER-1/ErbB1) is a member of the ErbB family of tyrosine kinase receptors (TK), which includes HER-1/ErbB1, HER-2/neu/ErbB2, HER-3/ErbB3, and HER-4/ErbB4. It is composed of extracellular (ligand binding), transmembrane and intracellular (tyrosine-kinase) domain. Upon ligand binding and receptor homo- or hetero-dimerization and activation, activated EGFR signals downstream to the PI3K/AKT and RAS/RAF/MAPK pathways (Fig. 10.1). These intracellular signaling pathways regulate key processes such as apoptosis and proliferation.

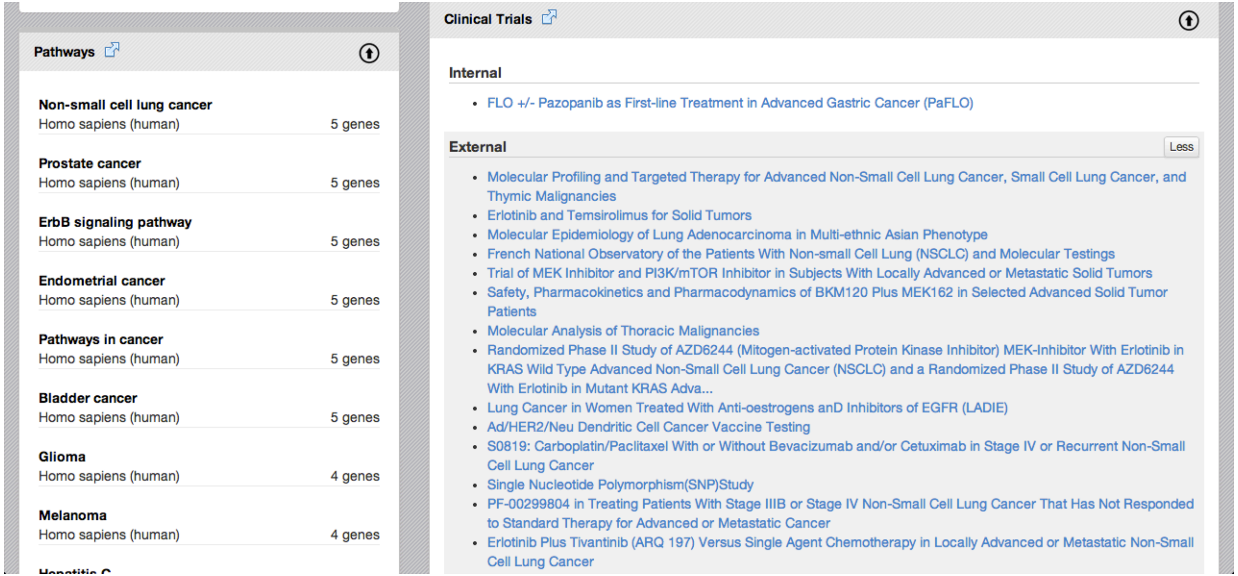
Therapies Targeting EGFR Pathway

Two treatment strategies targeting EGFR have been developed: monoclonal antibodies directed against the extracellular ligand-binding domain of the EGFR (e.g., cetuximab) and small molecule TK inhibitors (e.g., gefitinib [Iressa] and

- Interactively explore relevant publications, e.g. PDFs
- Improved ease of exploration, e.g. by highlighted medical terms and relevant concepts

Medical Knowledge Cockpit

Relevant Scientific Findings at a Glance



Pathways

Non-small cell lung cancer
Homo sapiens (human) 5 genes

Prostate cancer
Homo sapiens (human) 5 genes

ErbB signaling pathway
Homo sapiens (human) 5 genes

Endometrial cancer
Homo sapiens (human) 5 genes

Pathways in cancer
Homo sapiens (human) 5 genes

Bladder cancer
Homo sapiens (human) 5 genes

Glioma
Homo sapiens (human) 4 genes

Melanoma
Homo sapiens (human) 4 genes

Hepatitis C

Clinical Trials

Internal

- FLO +/- Pazopanib as First-line Treatment in Advanced Gastric Cancer (PaFLO)

External Less

- Molecular Profiling and Targeted Therapy for Advanced Non-Small Cell Lung Cancer, Small Cell Lung Cancer, and Thymic Malignancies
- Erlotinib and Temsirolimus for Solid Tumors
- Molecular Epidemiology of Lung Adenocarcinoma in Multi-ethnic Asian Phenotype
- French National Observatory of the Patients With Non-small Cell Lung (NSCLC) and Molecular Testings
- Trial of MEK Inhibitor and PI3K/mTOR Inhibitor in Subjects With Locally Advanced or Metastatic Solid Tumors
- Safety, Pharmacokinetics and Pharmacodynamics of BKM120 Plus MEK162 in Selected Advanced Solid Tumor Patients
- Molecular Analysis of Thoracic Malignancies
- Randomized Phase II Study of AZD6244 (Mitogen-activated Protein Kinase Inhibitor) MEK-Inhibitor With Erlotinib in KRAS Wild Type Advanced Non-Small Cell Lung Cancer (NSCLC) and a Randomized Phase II Study of AZD6244 With Erlotinib in Mutant KRAS Adva...
- Lung Cancer in Women Treated With Anti-oestrogens anD Inhibitors of EGFR (LADIE)
- Ad/HER2/Neu Dendritic Cell Cancer Vaccine Testing
- S0819: Carboplatin/Paclitaxel With or Without Bevacizumab and/or Cetuximab in Stage IV or Recurrent Non-Small Cell Lung Cancer
- Single Nucleotide Polymorphism(SNP)Study
- PF-00299804 in Treating Patients With Stage IIIB or Stage IV Non-Small Cell Lung Cancer That Has Not Responded to Standard Therapy for Advanced or Metastatic Cancer
- Erlotinib Plus Tivantinib (ARQ 197) Versus Single Agent Chemotherapy in Locally Advanced or Metastatic Non-Small Cell Lung Cancer

- Personalized clinical trials, e.g. by incorporating patient specifics
- Classification of internal/external trials based on treating institute