Building a Genomics Resource

From Experiments to APIs

Michael Baudis | UZH BIO390 HS20

progenetix

Genome screening at the core of “Personalised Health”

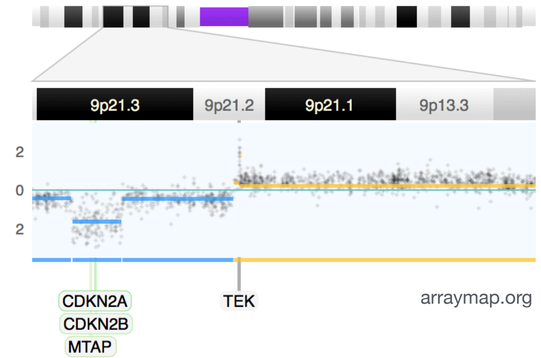
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**
- **Standardized data formats** and **exchange protocols** are needed to connect these resources throughout the world, for reciprocal, international **data sharing** and **biocuration** efforts
- Our work @ UZH:
  - ***cancer*** genome repositories
  - biocuration
  - protocols & formats

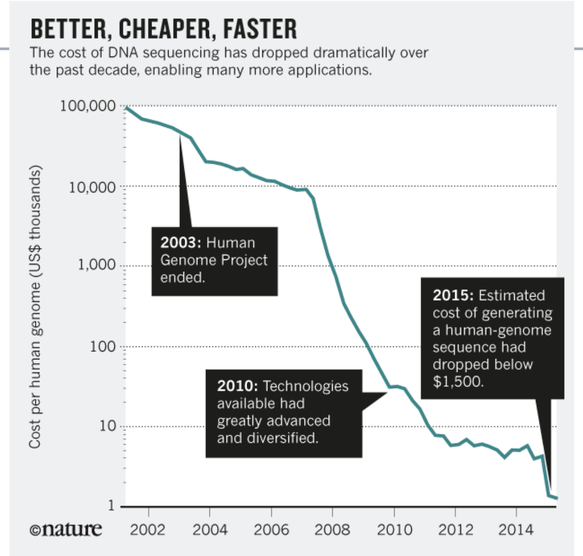
arrayMap

progenetix

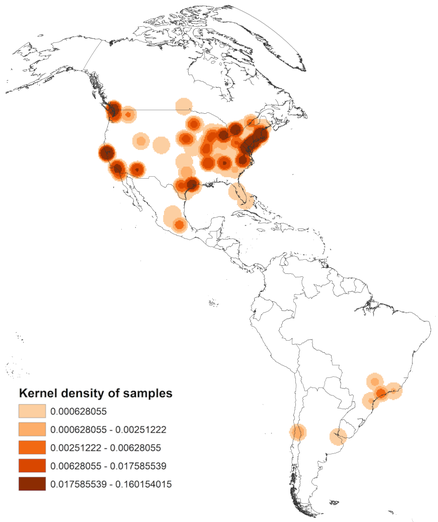
elixir

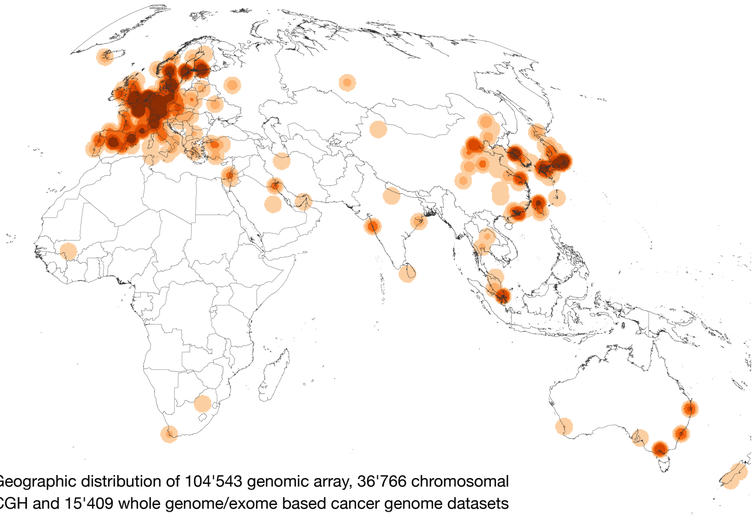
Global Alliance for Genomics & Health





The future of DNA sequencing. Eric D. Green, Edward M. Rubin & Maynard V. Olson. Nature; 11 October 2017 (News & Views)





Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

Building a Genomics Resource

A (personal) journey through time...

- Genomic Copy Number Variations in cancer (CNA / CNV)
- Comparative Genomic Hybridization (CGH) as the original CNV screening technique
- CNVs differ between cancer (sub)types and may correlate to clinical outcome
- single studies are limited in understanding disease-specific changes - let's build a database
- databases should be accessible - let's move online
- more data - data parsers & text mining
- visualization - graphics libraries and data formatting
- large datasets - access through APIs

progenetix

Theodor Boveri (1914)

Observations in sea urchin eggs

- **Cell-cycle checkpoints** (“Hemmungseinrichtung”)
- **Tumour-suppressor genes** (“Teilungshemmende Chromosomen”), which may be overcome by external signals, and can be eliminated during tumour progression
- **Oncogenes** (“Teilungsfoerdernde Chromosomen”) that become amplified (“im permanenten Übergewicht”)
- **Progression** (benign to malignant), w/ sequential changes of chromosomes
- Clonal origin & Genetic mosaicism
- Cancer **predisposition** through inheritance of “chromosomes” that are less able to suppress malignancy
- Inheritance of the same 'weak chromosome' from both parents leads to **homozygosity** and, consequently, to high-penetrance cancer syndromes - (e.g. xeroderma pigmentosum)
- Wounding and inflammation in tumour promotion; loss of cell adhesion in metastasis; sensitivity of malignant cells to radiation therapy (based on Hertwig *et al.*)

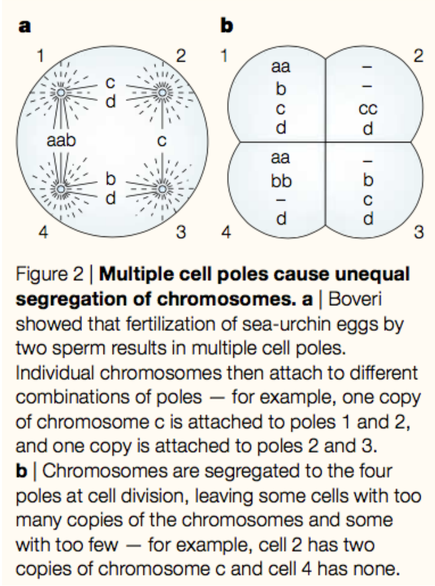


Figure 2 | **Multiple cell poles cause unequal segregation of chromosomes. a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached to poles 1 and 2, and one copy is attached to poles 2 and 3. **b** | Chromosomes are segregated to the four poles at cell division, leaving some cells with too many copies of the chromosomes and some with too few — for example, cell 2 has two copies of chromosome c and cell 4 has none.

Allan Balmain
Cancer genetics: from Boveri and Mendel to microarrays.
NatRev Cancer (2001); 1: 77-82

Anna Di Lonardo , Sergio Nasi , Simonetta Pulciani
Cancer: We Should Not Forget The Past
Journal of Cancer (2015), Vol. 6: 29-39
(for book cover & summary)

Janet Rowley (1972/73)

Chromosomal translocations in cancer

- Recurrent chromosomal translocations in leukemias and lymphomas
- "Philadelphia chromosome" in CML (Nowell & Hungerford, 1960) represents a reciprocal translocation between chromosomes 9 and 22
- 1972: t(8;21) ALL manuscript rejected by NEJM
- 1973: t(9;22) manuscript rejected by *Nature* "*with some reasonable comments and some truly wrong*"
- Clinical implications: **Tyrosine Kinase inhibitors** as standard first-line therapy in CML
  - first trials in 1998 (STI-571; Imatinib/Gleevec)
  - cf. Druker BJ, Lydon NB (2000). Lessons learned from the development of an Abl tyrosine kinase inhibitor… J Clin Invest 2000;105:3-7)

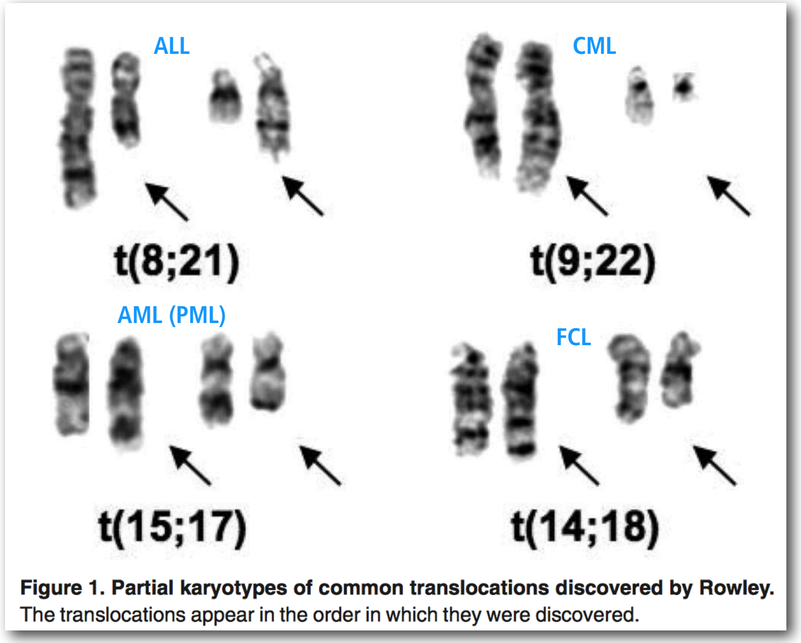
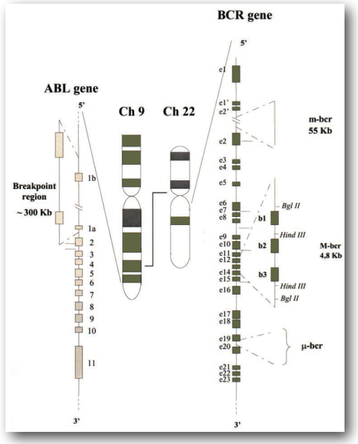
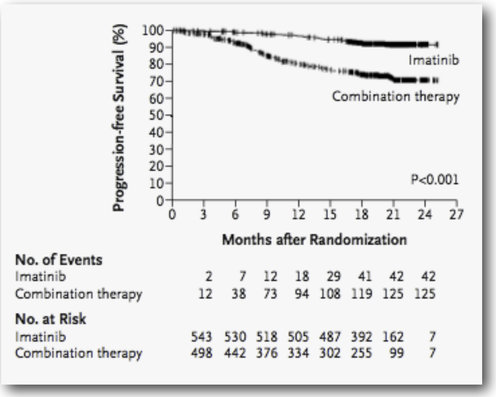


Figure 1. Partial karyotypes of common translocations discovered by Rowley. The translocations appear in the order in which they were discovered.

Janet D Rowley. Chromosomal translocations: revisited yet again Blood (2008), 112(6)



Pane et al. BCR/ABL genes …. Oncogene (2002), 21 (56)



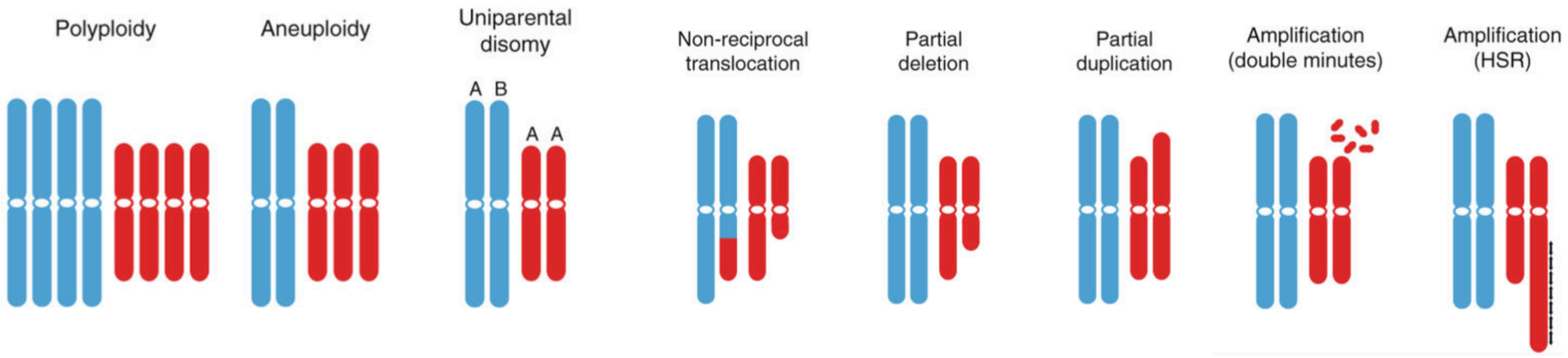
Event free Survival in first large Imatinib Trials

O'Brien et al. Imatinib compared with interferon and low-dose cytarabine… NEJM (2003) vol. 348 (11)

Types of genomic alterations in Cancer

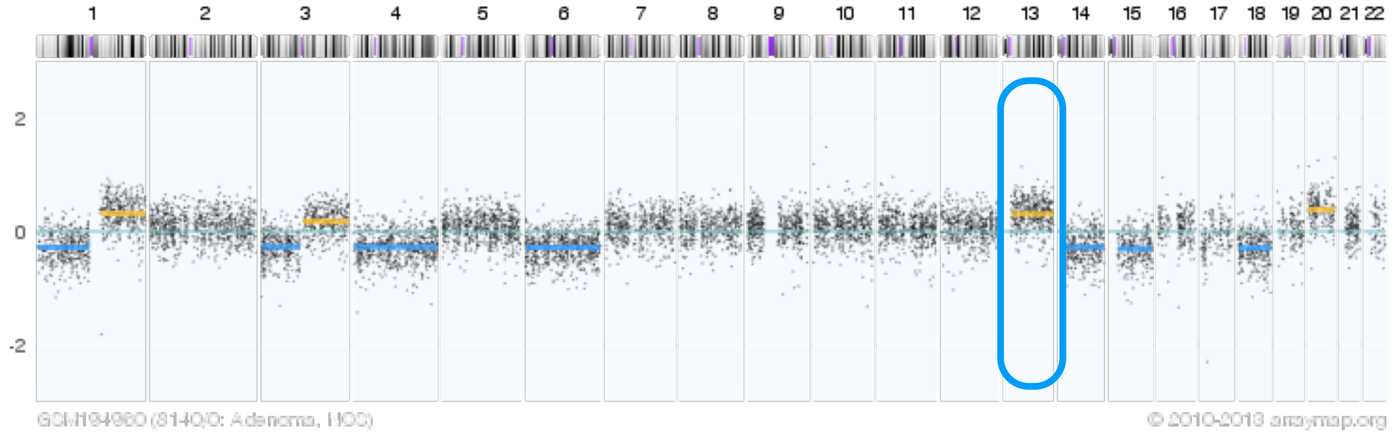
Imbalanced Chromosomal Changes: CNV

- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)

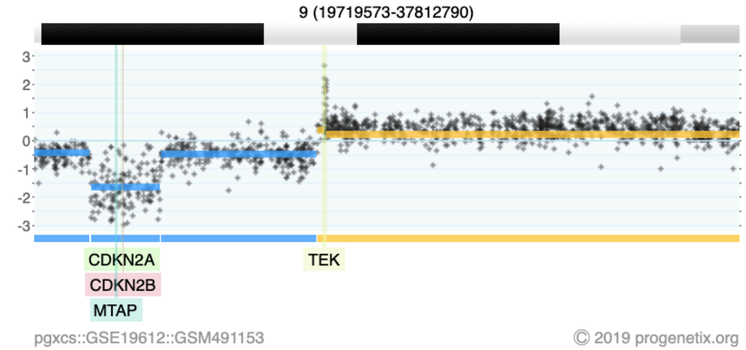


Grade et al., 2015 Recent Results Cancer Res

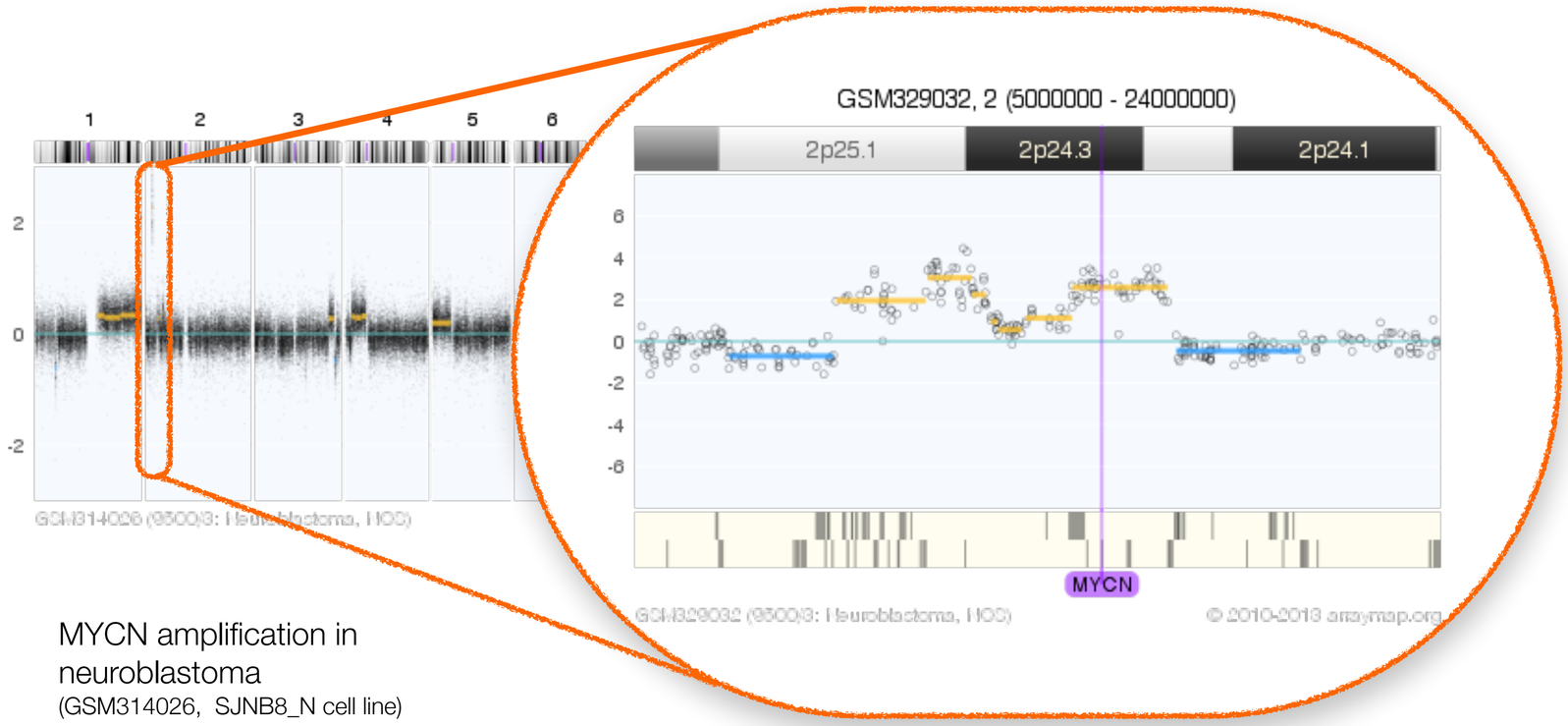
Somatic Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma

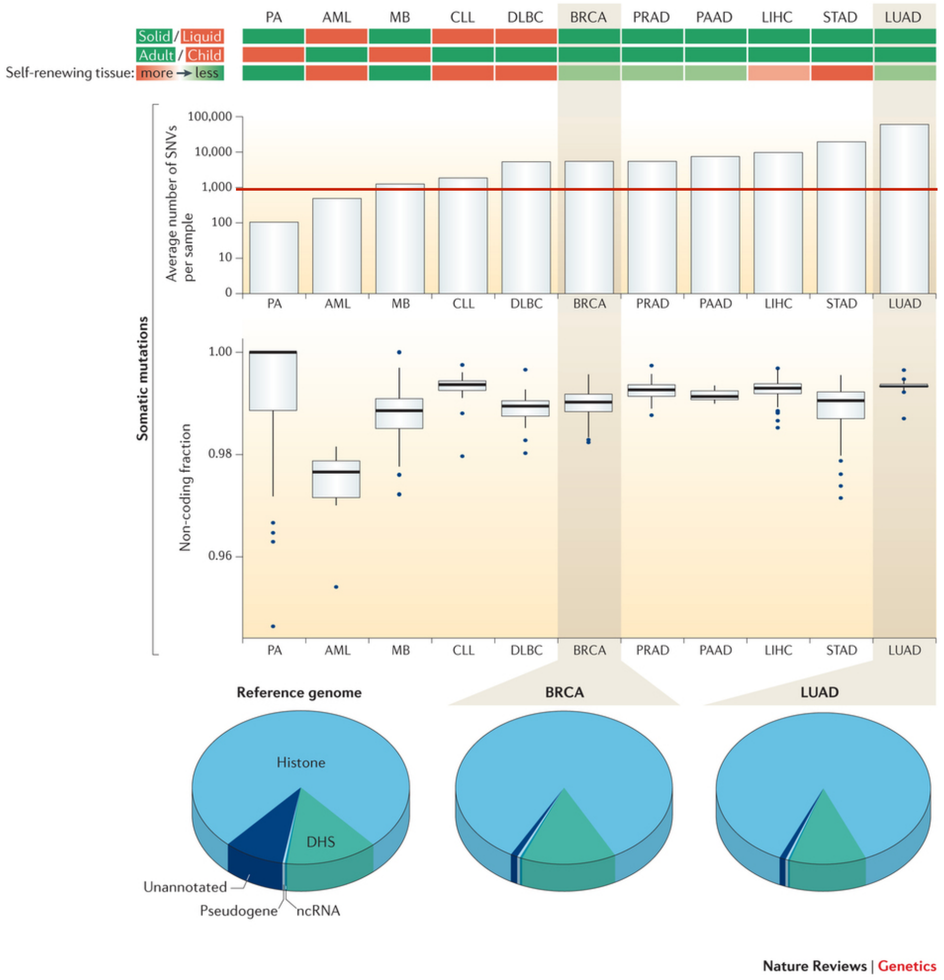


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

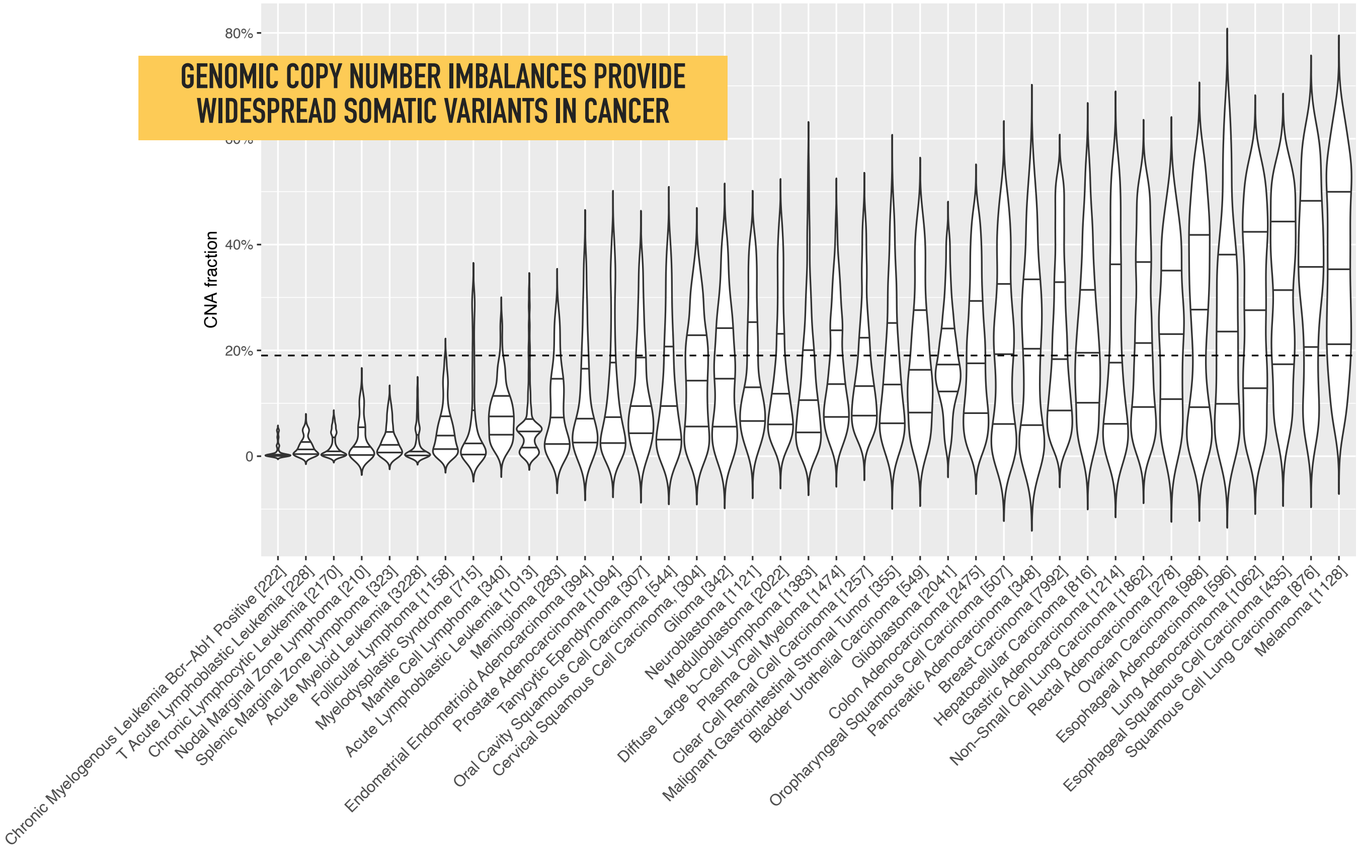
low level/**high level** copy number alterations (CNAs)

arrayMap

Quantifying Somatic Mutations In Cancer



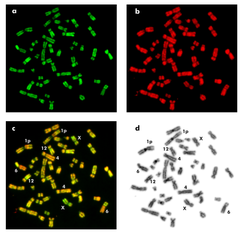
CANCERS SHOW THOUSANDS OF SINGLE NUCLEOTIDE VARIANTS PER SAMPLE, MOSTLY IN NON-CODING REGIONS



Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016)

On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from progenetix.org

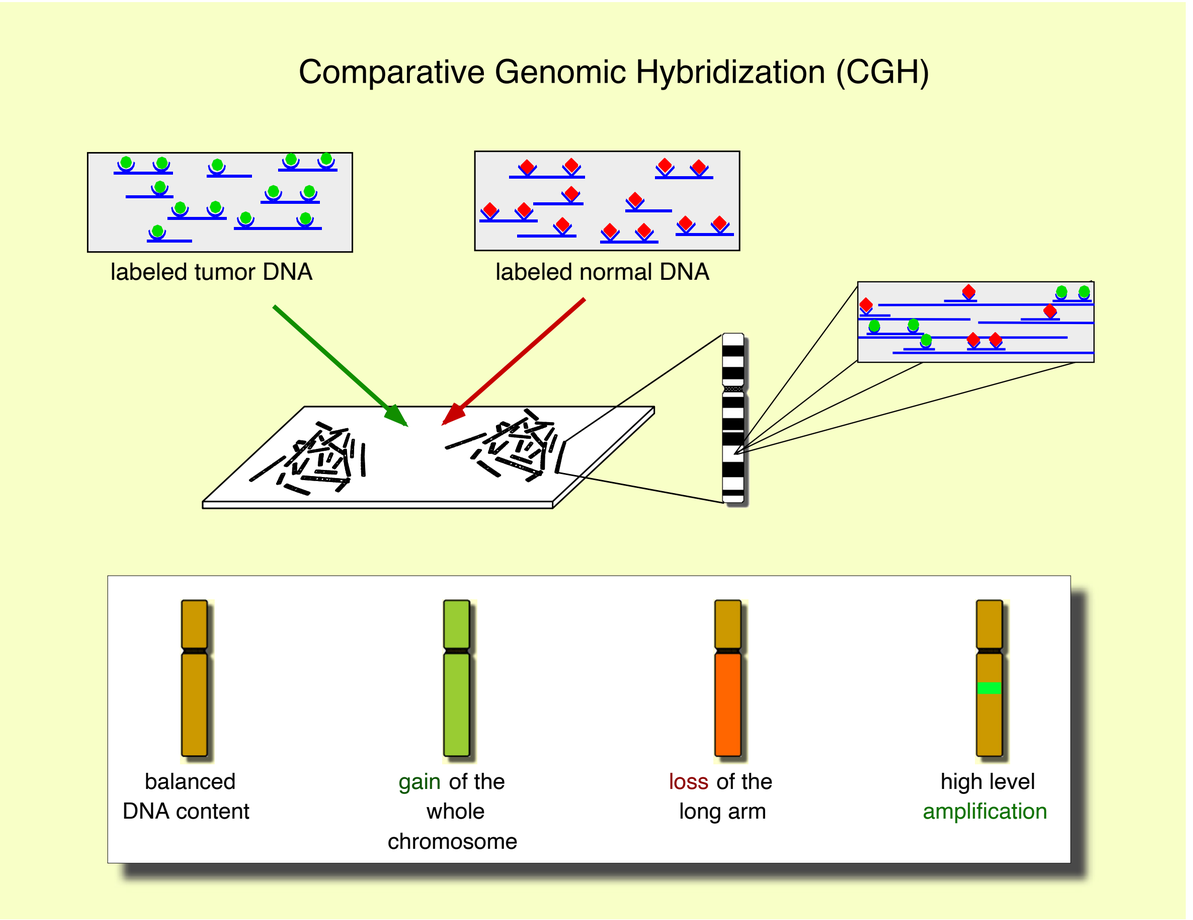
Comparative Genomic Hybridization



Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **semi-quantitative copy number** read-out
- **indirect** attribution of involved target genes through cytogenetic bands (megabase resolution)

- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;5083:818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. Hum Genet. 1993;90:584-589.



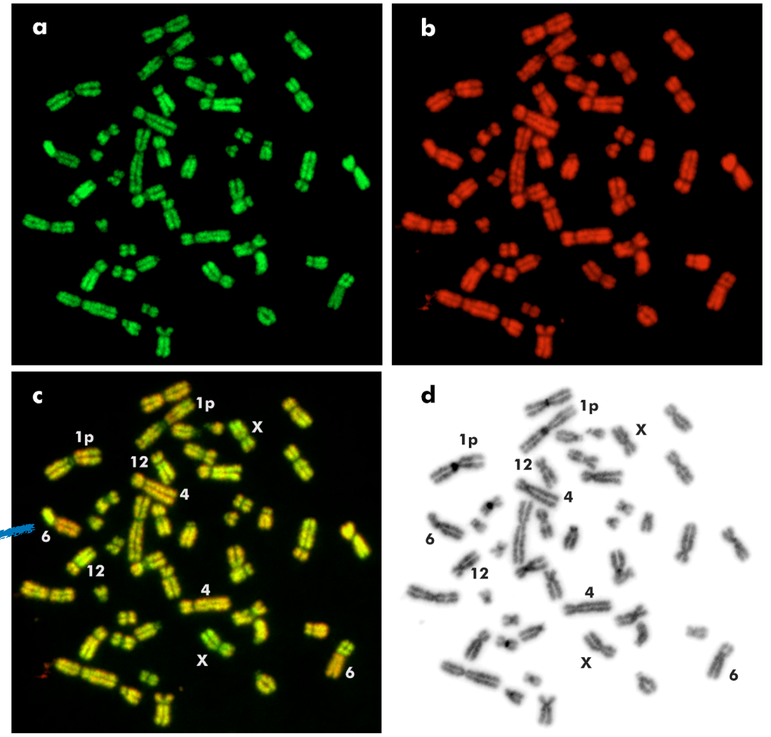
Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes in the copy number in the tumor DNA. Michael Baudis, 1998

Comparative Genomic Hybridization

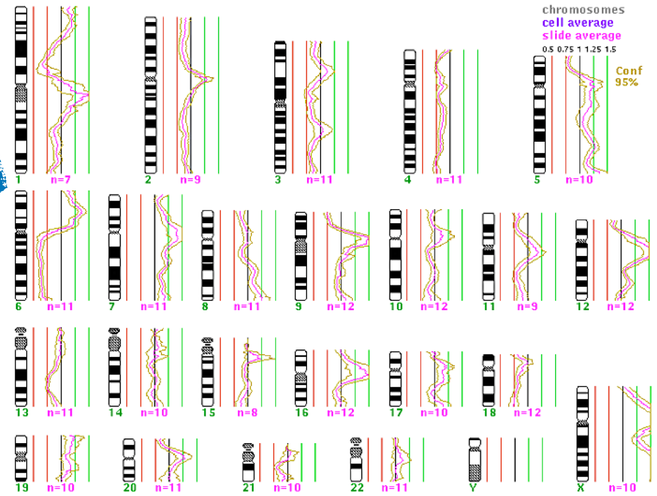
Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **semi-quantitative copy number** read-out
- **indirect** attribution of involved target genes through cytogenetic bands (megabase resolution)

- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;5083:818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. Hum Genet. 1993;90:584-589.



CGH-Experiment: **a** Hybridisierung mit Tumor-DNA; **b** Hybridisierung mit normaler menschlicher DNA als Kontrolle; **c** Überlagerung der Signale; **d** Bänderungsfärbung zur Identifizierung der Chromosomen



Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

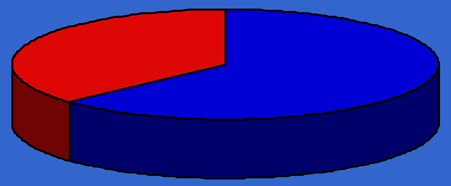
Cancer CNVs | Diagnostics | Prognosis

Single-study CNV frequencies correspond to diagnostic subsets

Einleitung
CGH
Diagramm
Häufigste Veränderungen
Histologien
cb vs. ib
Zusammenfassung
Dank

Molekularzytogenetische Analyse agressiver non-Hodgkin Lymphome im Rahmen der NHL-B Studie der DSHNHL

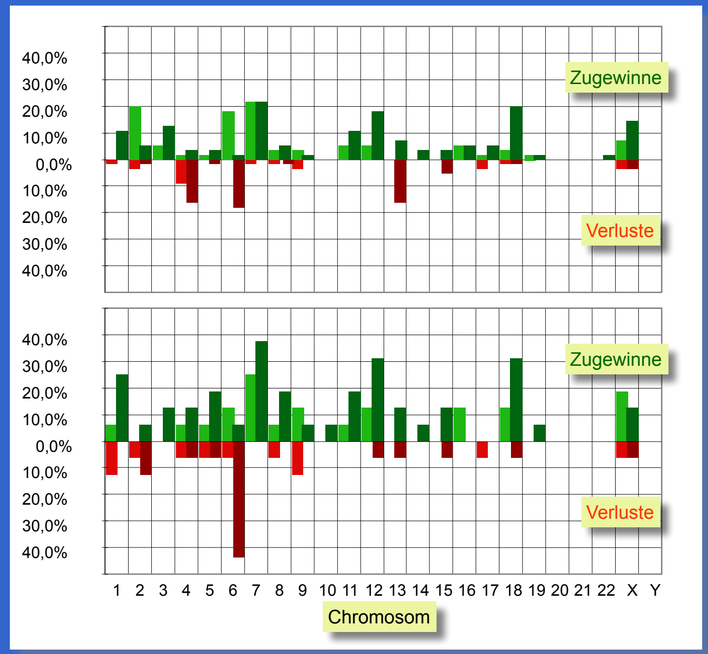
- **CGH (comparative genomic hybridization):**
  -Untersuchung des gesamten Genoms in einem einzigen Experiment unter Verwendung genomischer DNA
  -Verwendbarkeit von Paraffinmaterial
- **144 der 228 Gewebeproben primärer Studienpatienten bisher auswertbar**



55 cb NHL

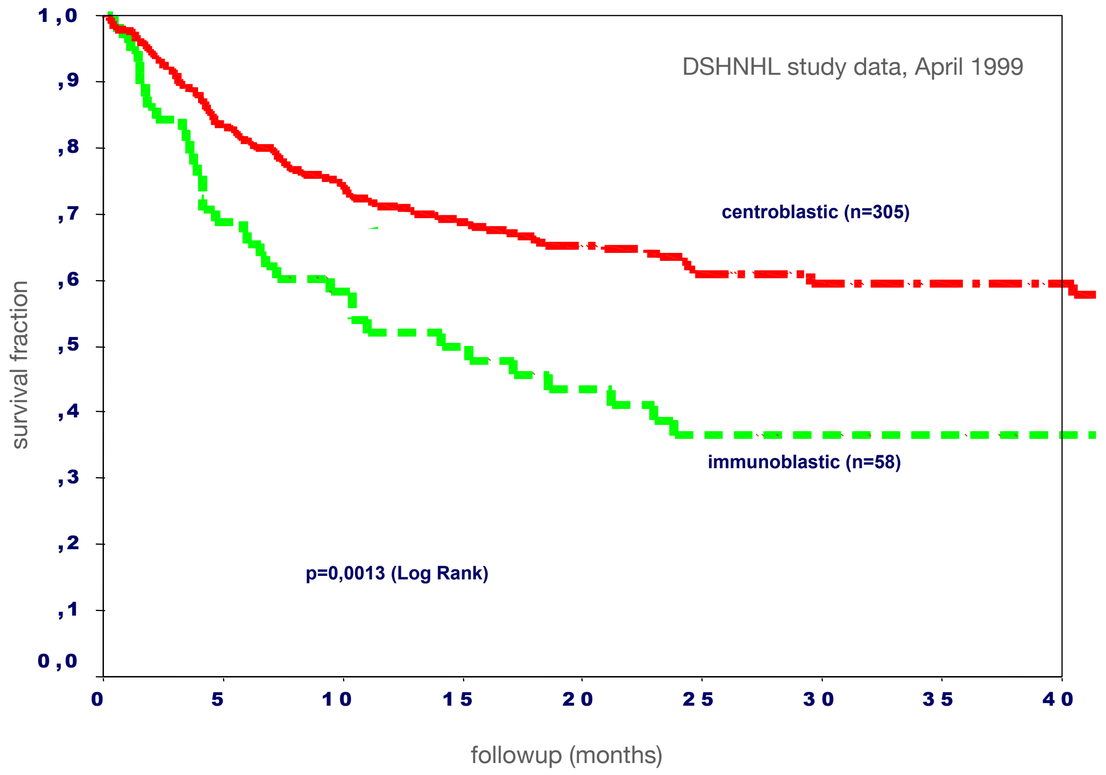
Chromosomale Imbalancen

16 ib NHL



Cancer CNVs | Diagnostics | Prognosis

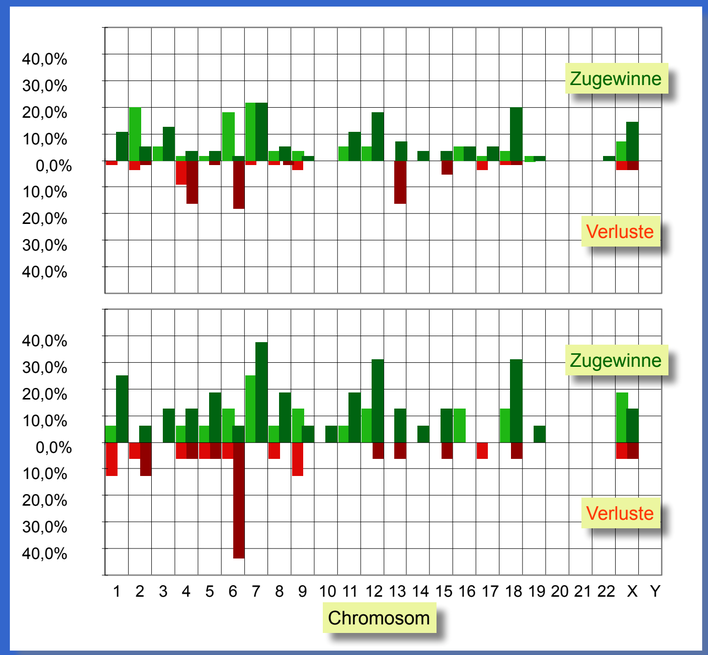
Single-study CNV frequencies correspond to diagnostic subsets



55 cb NHL

Chromosomale Imbalancen

16 ib NHL



dilbert.com | Tuesday February 27, 1996

Let's build a database!

... using archaic tools

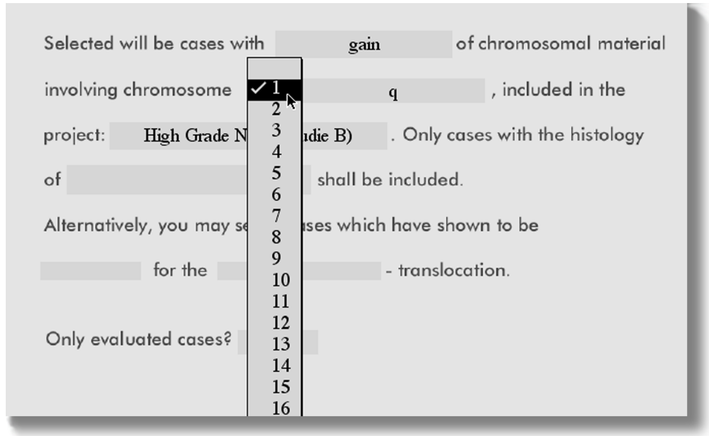
dilbert.com | Tuesday September 08, 1992

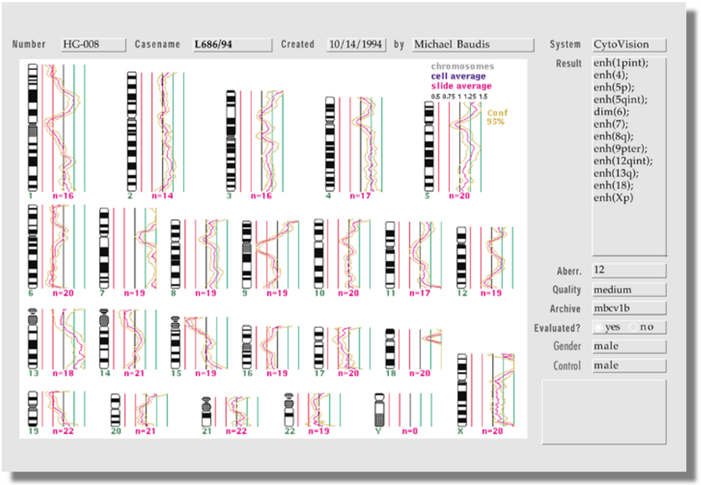
Progenetix CGH Database and Website

HEIDELBERG UNIVERSITY HOSPITAL

dkfz.

- originally an internal FileMaker Pro database, to store CGH profiles and annotations for the "Organization of Complex Genomes" group (head: Peter Lichter) at the German Cancer Research Center (DKFZ), starting in 1998
- expansion to include literature derived data, with a focus on malignant non-Hodgkin's lymphomas
- in 2000 online version






```
Domain Name: PROGENETIX.NET
Registry Domain ID: 45628826_DOMAIN_NET-VRSN
Registrar WHOIS Server: whois.enterprice.net
Registrar URL: http://www.epag.de
Updated Date: 2019-06-01T04:20:49Z
Creation Date: 2000-11-29T18:17:38Z
```

- Dec 6, 2000
  - first time online
- Nov 30, 2000
  - addition of graphical representation and gene table
- Nov 17, 2000
  - generation of website layout and database automatisation

progenetix.net: storage and visualization of genomic aberration data in human malignancies

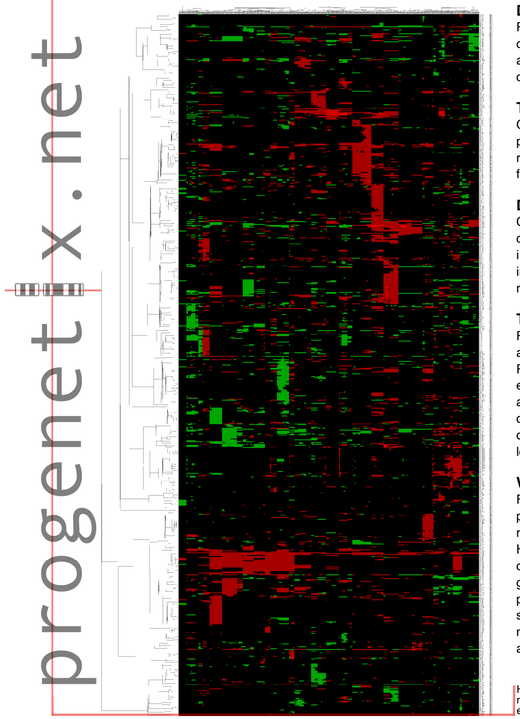
michael baudis, md

Over the last decade, techniques for the genome wide scanning for genomic imbalances in malignant neoplasia have been developed, e.g. Comparative Genomic Hybridization (CGH).

Currently, no comprehensive online source for CGH data with a standardized format suitable for data mining procedures has been made available for public access. Such a data repository could be valuable in identifying genetic aberration patterns with linkage to specific disease entities, and provide additional information for validating data from large scale expression array experiments.

A case and band specific aberration matrix was selected as most suitable format for the mining of CGH data. The [progenetix.net] data repository was developed to provide the according data to the research community for a growing number of human malignancies.

In the current implementation, two main purposes are being served. First, access to the band specific pattern of chromosomal imbalances allows the instantaneous identification of genomic “hotspots”. Second, the band specific aberration matrices can be included in data mining efforts. As an example, the clustering off all informative cases from the current (September 2001) dataset is shown here (online source under www.progenetix.net/bcats/clustered.png).



Hierarchical clustering of band specific chromosomal imbalances from 999 human neoplasias, contained in the [progenetix.net] collection. Cases without aberrations were excluded.

Data selection

PubMed is searched for publications applying CGH to the analysis of malignant tumors. Articles are selected according to their online availability and the description of genomic imbalances on a per case basis.

Transformation of input data

Chromosomal aberration data is transformed via customized parsing commands to a common format adherent to ISCN 1995 recommendations. In some cases, aberration data was transcribed from graphical representations or provided by the authors.

Data storage

Currently, the primary data is stored in a dedicated “off-line” database. Besides case identifier and ISCN adapted chromosomal imbalance data, tumor classification and source information including the PubMed identifier is recorded. Disease entities are reclassified to ICD-O-3 codes.

Text parsing and generation of aberration matrix

For the generation of the case and band specific aberration matrix, a dedicated text pattern comparison model was developed using Perl. Briefly, for each chromosomal band, the aberration field of each case is searched for a variety of patterns containing aberration information applying to that band. A matrix with currently 324 band resolution is generated, annotating chromosomal gains with “1” and losses with “–1”; localized high-level gains are designated “2”.

Website generation

For graphical representation of chromosomal imbalances, HTML pages containing different views of the underlying aberration matrices are generated using Perl. Graphics are implemented using HTML syntax. Besides band specific, whole genomic overviews, chromosome specific pages with links to all involved cases are generated for each ICD-O-3 entity as well as for each registered project. Additionally, those representations are available for several subsets combining related data (e.g. all lymphoid neoplasias, breast carcinoma cases). For each of the groups, the according aberration matrix is linked for download.

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 17 no. 12 2001*
Pages 1228–1229

Progenetix.net: an online repository for molecular cytogenetic aberration data

Michael Baudis [1, 2,] and Michael L. Cleary [2]*

[1]Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany and [2]Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA

Received on July 5, 2001; revised on July 9, 2001; accepted on July 16, 2001

ABSTRACT

Summary: Through sequencing projects and, more recently, array-based expression analysis experiments, a wealth of genetic data has become accessible via online resources. In contrast, few of the (molecular-) cytogenetic aberration data collected in the last decades are available in a format suitable for data mining procedures. www.progenetix.net is a new online repository for previously published chromosomal aberration data, allowing the addition of band-specific information about chromosomal imbalances to oncologic data analysis efforts.

Availability: http://www.progenetix.net

Contact: mbaudis@stanford.edu

Neoplastic transformation and progression is the result of genetic defects arising in normal cells and giving rise to a malignant clone. During the process of oncogenesis, some of the usually multiple steps required for acquisition of the full neoplastic phenotype may represent themselves as numerical or structural abnormalities in the chromosomes of the transformed cells.

Over the last decades, the analysis of chromosomal abnormalities in malignant cells has gained importance in oncologic research as well as in clinical practice. A vast number of genetic abnormalities has been identified in the virtually complete range of human neoplasias. Several attempts have been undertaken for collection and classification of those abnormalities, the most widely recognized being the catalog by Mitelman and co-workers (Mitelman, 1994; online access through http://cgap.nci.nih.gov/Chromosomes/Mitelman).

In addition to metaphase analysis of short-term cultivated tumor cells or tumor cell lines, molecular cytogenetic techniques have recently been applied to the analysis of chromosomal abnormalities in primary tumor tissues. One of the more widely used screening techniques is Comparative Genomic Hybridization (CGH; Kallioniemi *et al.*, 1992; du Manoir *et al.*, 1993). Briefly, this method is based on the competitive *in-situ* hybridization of differentially labeled tumor versus normal genomic DNA to normal human metaphase spreads. The calculation of the intensity ratios of the two fluorochromes gives an overview about relative gains and losses of DNA in the tumor genome with mapping to the respective chromosomal bands. The identification of frequently imbalanced regions in tumor entities may point towards tumor suppressor gene or proto-oncogenes mapping to the respective chromosomal bands. Usually, the result of those experiments is communicated either in text format according to the International System for Cytogenetic Nomenclature (Mitelman, 1995) or graphically, with aberration bars next to chromosomal ideograms for the representation of chromosomal gains and losses.

Because in each experiment CGH analysis covers the whole number of chromosomes, the comparision of data sets from related malignancies could lead to the delineation of common as well as divergent genetic pathways defining the respective malignant phenotypes. Although an extremely large number of malignant tumors has been analyzed using this technique, no comprehensive CGH database with band-specific chromosomal aberration information is publicly available[†].

A minimal requirement for such a database would be the conversion of the text or graphical information used in publications to data tables, representing the information about the aberration status of single chromosomal bands for each case. For the site discussed here, this process includes: (1) the transformation of the published results in a format adapted from the ISCN, and (2) the automatic generation of the band specific aberration table.

Due to format variations of the published data, step 1 consists of the manual conversion of the text data or evaluation and conversion of the graphical representations, respectively. Due to the (in computational terms) odd

*To whom correspondence should be addressed.

[†]Links to a number of online CGH resources with different scopes can be found at www.progenetix.net.

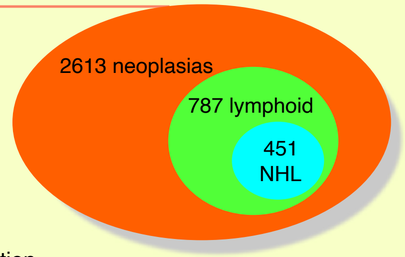
progenetix.net

[progenetix.net] online CGH database

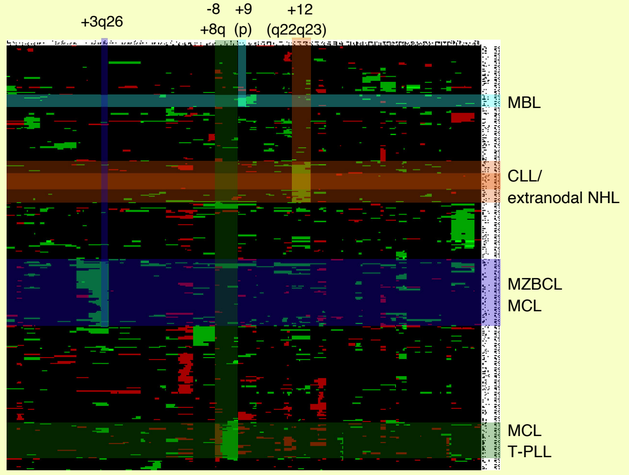
A compilation of published CGH analyses with reported case and band specific results

Currently includes
2613 cases from
92 publications

Automatic
conversion of
ISCN format to
aberration matrix
with 393 bands resolution



Clustering in 451 NHL: +12 in CLL and extranodal NHL



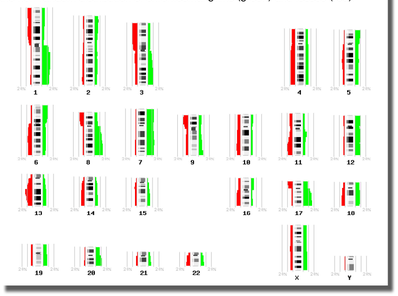
progenetix.net

Collection and Transformation of Chromosomal Imbalances in Human Neoplasias for Data Mining Procedures

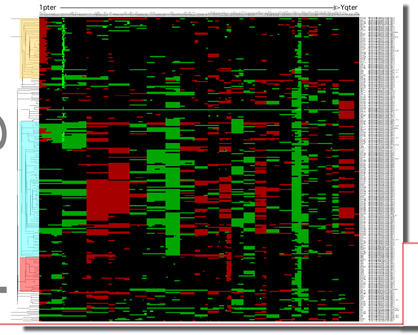
michael baudis, dept. of pathology, stanford university

Although the deciphering of the human genome has been pushed forward over the last years, little effort has been made to collect and integrate the treasure trove of clinical tumor cases analyzed by molecular-cytogenetic methods into current data schemes. Publicly announced at BCATS 2001, since then [progenetix.net] has been established as the largest public source of chromosomal imbalance data with band-specific resolution. Targets for the use of the data collection may be the description of prediction of oncogene and suppressor gene loci, identification of related loci for pathway creation, and especially the combination of the data with expression array experiments for filtering of relevant genes among the deregulated candidates.

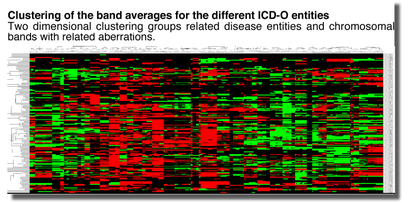
Chromosomal imbalances in 5478 clinical cases from 196 publications
Although not as prominent as in specific subgroups, this large collection shows the non-random distribution of chromosmal gains (green) and losses (red).



Material and Methods Chromosomal aberration data of more than 5478 cases from 196 publications describing results of Comparative Genomic Hybridization (CGH) experiments were collected. Minimal requirements were diagnosis of a malignant or benign neoplasia, analysis of clinical tumor samples and report of the analysis results on a case by case basis, resolved to the level of single chromosomal bands.Data was transformed from the diverse annotation formats to standardized ISCN “rev ish” nomenclature. For the transformation of the non-linear ISCN data to a two-dimensional matrix with code for the aberration status of each chromosomal band per case, a reverse pattern matching algorithm was developed in Perl. Graphical representations and cluster images are generated for all different subsets (Publications, ICD-O-3 entities, meta-groups) and presented on the progenetix.net website.

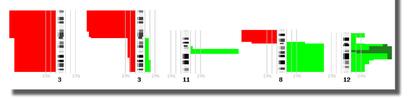


Clustering of the band averages for the different ICD-O entities
Two dimensional clustering groups related disease entities and chromosomal bands with related aberrations.



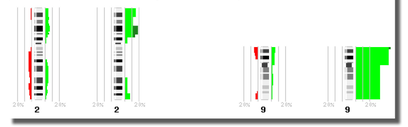
Results Out of 4896 tumor samples, 3862 (79%) showed chromosomal imbalances by CGH. The average per band probability was 4.5% for a loss (max. 12.9% at 13q21) and 6.5% for a gain (max. 15.6% at 8q23). Differences between neoplastic entities showed in the average frequency and distribution pattern of imbalanced chromosomal regions. Tumor subsets (10 or more cases) with the strongest hot spots for losses were small cell lung carcinomas (ave. 23.3% with max. 96.2% at 3p14p26) and pheochromocytomas (ave. 10.9% with max. 92.7% at 3p); prominent gain maxima were found in pure high grade infiltrating duct carcinomas of the breast (ave. 5.9% with max. 95.7% at 11q13), T-PLL (ave. 4.7% with max. 81.8% for whole 8q) and dedifferentiated liposarcomas (ave. 10.4% with max. 81.8% at 12q13), among others.
By cluster analysis, different combinations of chromosomal hot spot regions could be shown to occur in tumors subsummized in the same diagnostic entity; the example of neuroblastomas is shown.

Examples of hotspots of genomic imbalance
SCLC, pheochromocytoma, high grade DCIS, T-PLL, dedifferentiated liposarcoma



Conclusion So far, progenetix.net project was able to:
1. collect a large dataset of genomic aberration data generated through a molecular-cytogenetic screening technique (CGH)
2. develop the software tools to transform those data to a meta format compatible to commonly used genomic interval descriptions
3. produce graphical and numerical output from those data for hot spot detection and statistical analysis.
For future approaches, the data collection will be valuable for filtering data from expression array experiments for relevant genes, and possibly for the description of common and divergent genetic pathways in the oncogenetic process of different tumor entities.The transformed raw data of the progenetix.net collection is avaible for research purposes over the website.

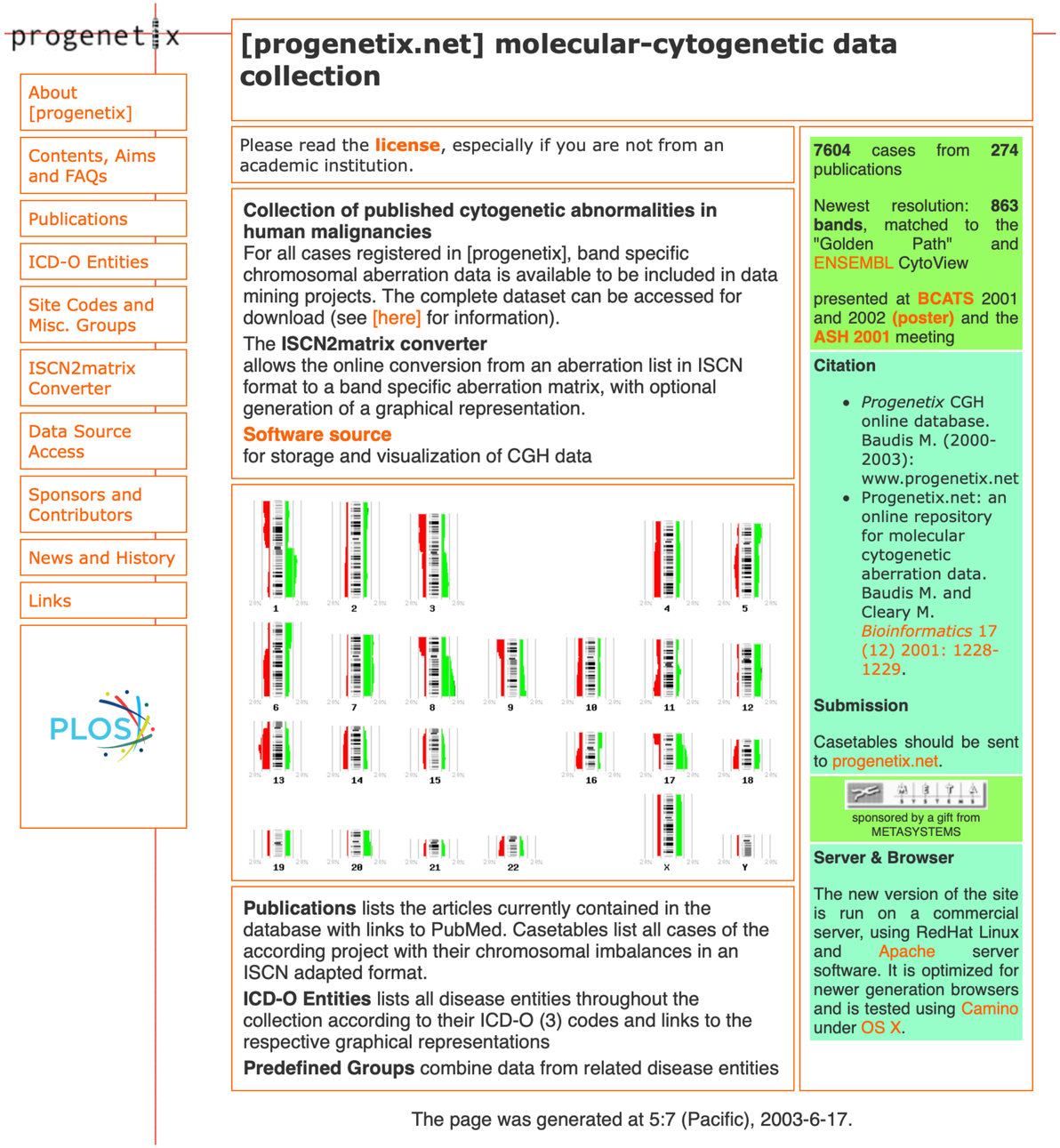
Distinction of histologically related through their chromosomal aberration pattern
Amplification of the REL locus on 2p16 and gain of 9p(ter) distinguishes primary mediastinal B-cell lymphomas (PMBL, right) from diffuse large cell lymphomas (DLCL, left). The distinction may have clinical implications



Identification of different aberration patterns in Neuroblastoma (289 cases)
N-Myc (2p25) amplification is the hallmark of a subgroup, showing only consistant loss of the terminal portion 1p. Other groups are defined by the loss of parts of 11q, or a “chromosomal instability” phenotype. Gains on 17q are a common feature of all groups.
Those patterns may be combined with gene-level information to reconstruct the different pathways leading to malignant transformation.

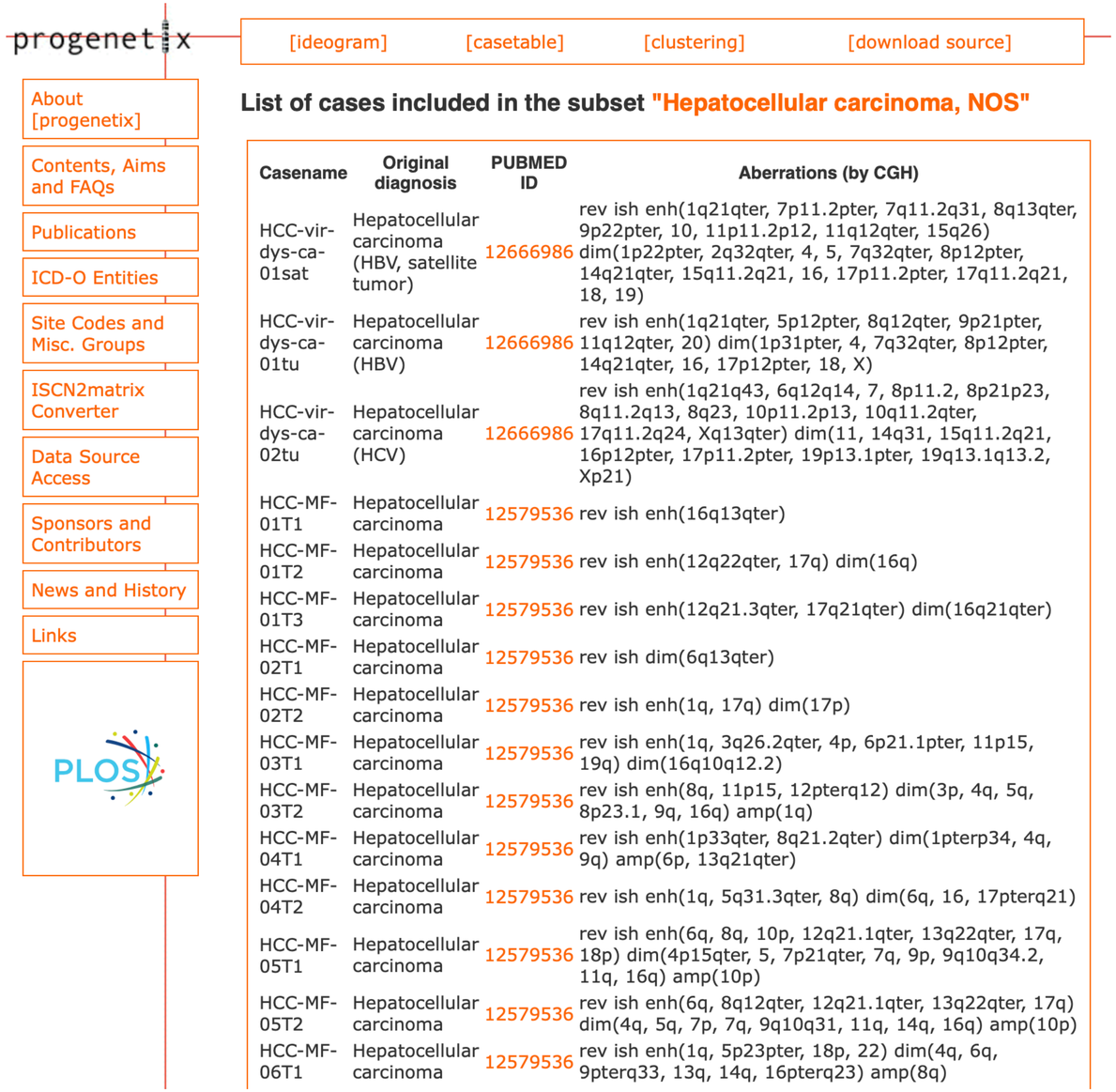
Progenetix Database in 2003
Text conversion for CNVs

- based on listed CGH results from publications
  - literature detection using optimized PubMed queries
  - extraction (copy/paste, typing) of rev ish ISCN karyotypes from articles and supplementary material
  - annotation cleanup using scripting with regular expressions (Perl)
  - custom script to convert cleaned ISCN annotations to cytoband status maps
  - custom graphics libraries to create graphical representations of CNV frequencies



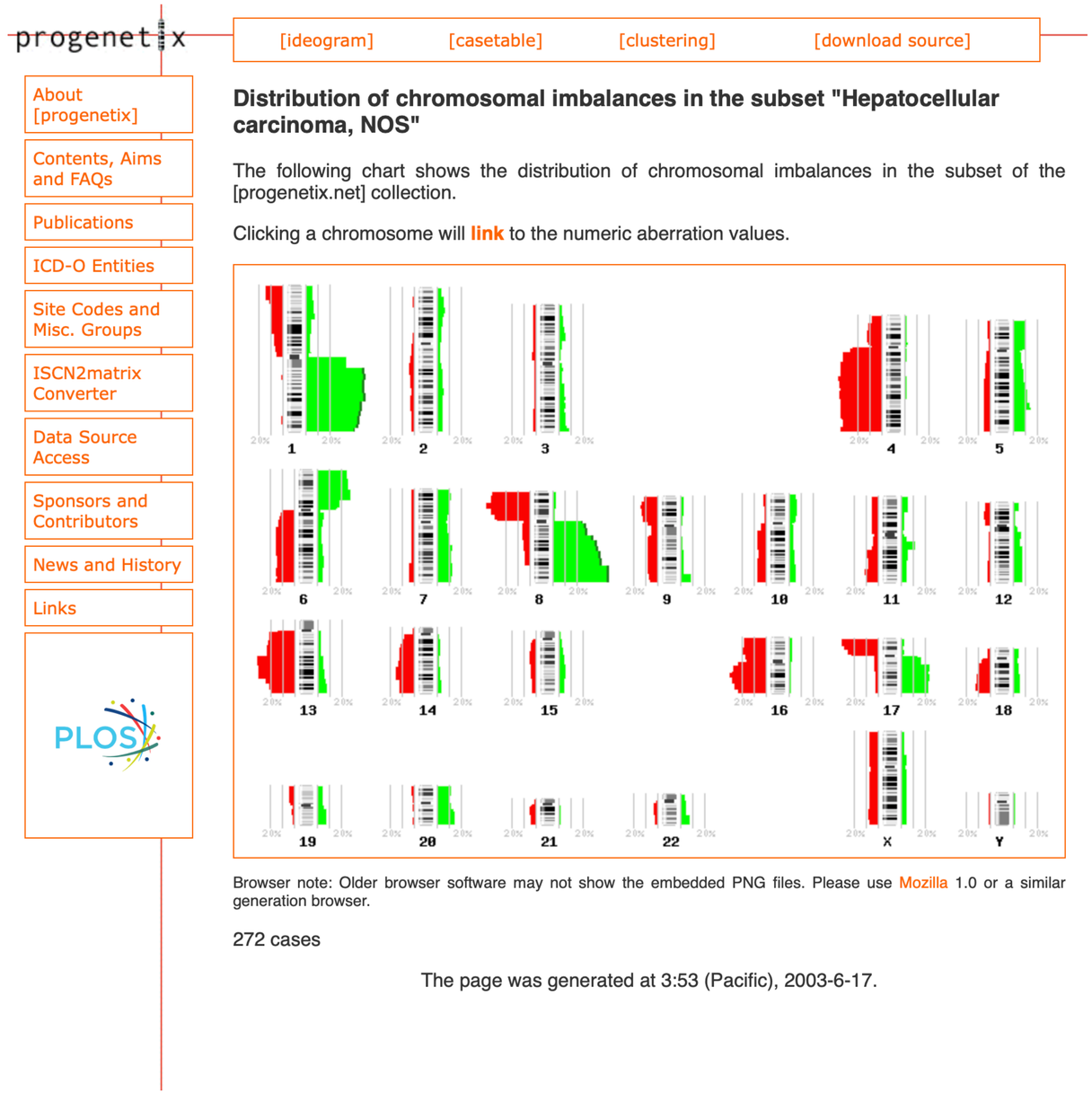
Progenetix Database in 2003
Text conversion for CNVs

- based on listed CGH results from publications
  - literature detection using optimized PubMed queries
  - extraction (copy/paste, typing) of rev ish ISCN karyotypes from articles and supplementary material
  - annotation cleanup using scripting with regular expressions (Perl)
  - custom script to convert cleaned ISCN annotations to cytoband status maps
  - custom graphics libraries to create graphical representations of CNV frequencies



Progenetix Database in 2003
Text conversion for CNVs

- based on listed CGH results from publications
  - literature detection using optimized PubMed queries
  - extraction (copy/paste, typing) of rev ish ISCN karyotypes from articles and supplementary material
  - annotation cleanup using scripting with regular expressions (Perl)
  - custom script to convert cleaned ISCN annotations to cytoband status maps
  - custom graphics libraries to create graphical representations of CNV frequencies



Progenetix Database in 2003
Text conversion for CNVs

- articles and supplements with **cytoband-based** *rev ish* CGH results
- sometimes rich, but **unstructured** associated information
- **PDFs** readable, but **not well suited for data** extraction (character entities, text flow)

progenetix

TABLE 3. Comparison of Primary Tumors and Metastases by CGH

Case	Gain in common	Loss in common	Primary tumor only	Metastasis only
108		18		
113	7, 8q24-qter, 13q11-qter, 20q11-qter, Xq11-Xter	1p33-pter, 2p21-pter, 4q24-qter, 15q11-q15, 17p11-pter, 18		
LM	12q22-qter, 15q23-qter, 17q11-ter, 20p11-p12, 20q11-ter, 22q11-ter	1p11-p32, 1q24-31, 4, 13q11-qter, 17p11-pter, 18, 20p11-ter	11p11-pter-	12+
145	4q26-q28, 6p11-p13, 8p11-p12, 920q11-qter	1p11-pter, 4q31-qter, 6q11-qter, 8p12-pter, 11, 15q11-qter, 16q11-qter, 17p11-pter, 18, 21q11-qter	13q21-qter+, 20p11-pter-	8q11-qter+,10-, 6p21-pter-
53	7, 8q11-qter, 9q33-qter, 13q11-qter, 20p11-p12, 20q11-qter	4p13-pter, 4q21-qter, 8p12-pter, 15q14-qter, 18q11-qter, 20p12-pter	5p11-pter-, 5q13-qter-, 14q11-qter-	11+, 16p11-pter+, 17q11-qter+, 19+, 21q11-qter+, 22q11-qter+
147	7, 13q11-qter, 20q11-qter	8p21-pter, 18	4p14-pter-, 4q28-qter+, 8p11-21-, 17q11-q2+, 21q11-qter-	11q22-qter+, 16+, 1p11-33-

TABLE 1. Clinical Data

Case number	Age	Sex	Site	Stage[a]	Grade[b]	Diagnosis of metastatic disease[c]
2	40	M	Transverse colon	IV	3	Synchronous
6	79	M	Ascending colon	IV	2	Synchronous
9	73	M	Transverse colon	II	2	N/A
11	56	M	Rectosigmoid	IV	2	Metachronous
12	70	F	Sigmoid colon	IV	2	Synchronous
13	65	M	Descending colon	II	9	Synchronous
14	60	M	Rectum	III	3	Metachronous
15	51	F	Rectum	III	2	Metachronous
19	63	M	Rectosigmoid Junction	III	2	Synchronous
20	63	M	Rectum	IV	9	Metachronous
21	64	F	Sigmoid colon	IV	2	Synchronous
35	71	M	Rectum	III	9	Metachronous
49	72	M	Cecum	IV	3	Synchronous
53	72	F	Sigmoid colon	IV	2	Synchronous
104	61	M	Sigmoid colon	IV	2	Metachronous
105	58	M	Ascending colon	II	2	Metachronous
107	77	F	Cecum	IV	2	Metachronous
108	53	F	Splenic flexure	IV	2	Synchronous
112	68	M	Rectum	III	3	Synchronous
113	41	M	Splenic flexure	IV	2	Synchronous
114	49	M	Splenic flexure	IV	3	Synchronous
116	73	M	Rectosigmoid	III	9	Metachronous
120	24	F	Descending colon	IV	2	Synchronous
123	62	F	Rectum	III	2	Metachronous
124	42	M	Rectum	IV	9	Synchronous
145	70	M	Rectosigmoid	IV	2	Synchronous
147	86	F	Cecum	IV	2	Synchronous

[a]AJCC/UICC staging system (Hutter and Sobin, 1986).
[b]Grade of primary tumor: 1–3, low, moderate, high grade; 9, grading unknown.
[c]Synchronous, diagnosis of metastatic disease within 12 months following diagnosis of primary tumor; metachronous, diagnosis of metastatic disease after 12 months or later.

GENES, CHROMOSOMES & CANCER 25:82-90 (1999)

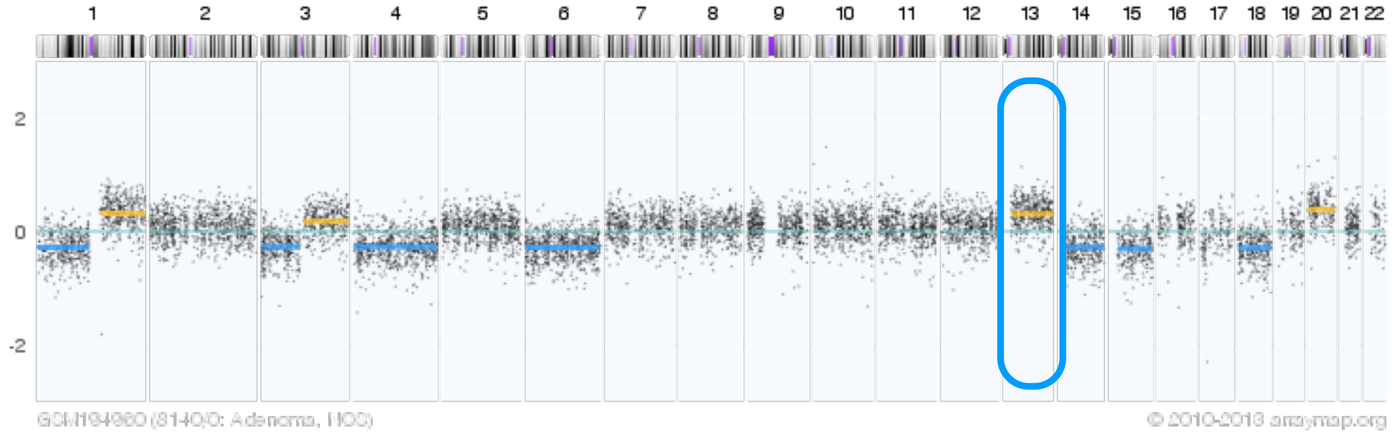
Chromosome Arm 20q Gains and Other Genomic Alterations in Colorectal Cancer Metastatic to Liver, as Analyzed by Comparative Genomic Hybridization and Fluorescence In Situ Hybridization

W. Michael Korn,[1]* Toru Yasutake,[2] Wen-Lin Kuo,[1] Robert S. Warren,[3] Colin Collins,[1] Masao Tomita,[2] Joe Gray,[1] and Frederic M. Waldman[1]

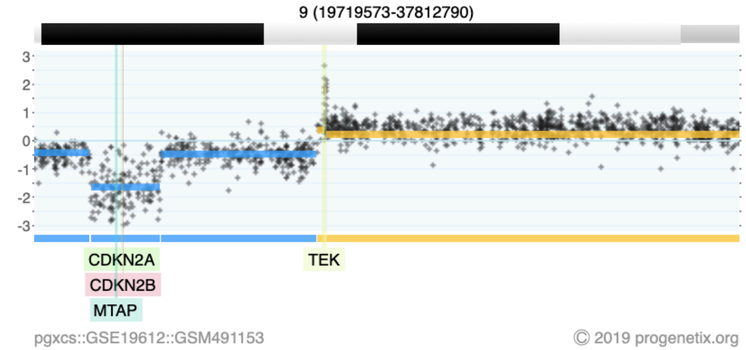
https://progenetix.org/2003-06-17/

progenetix

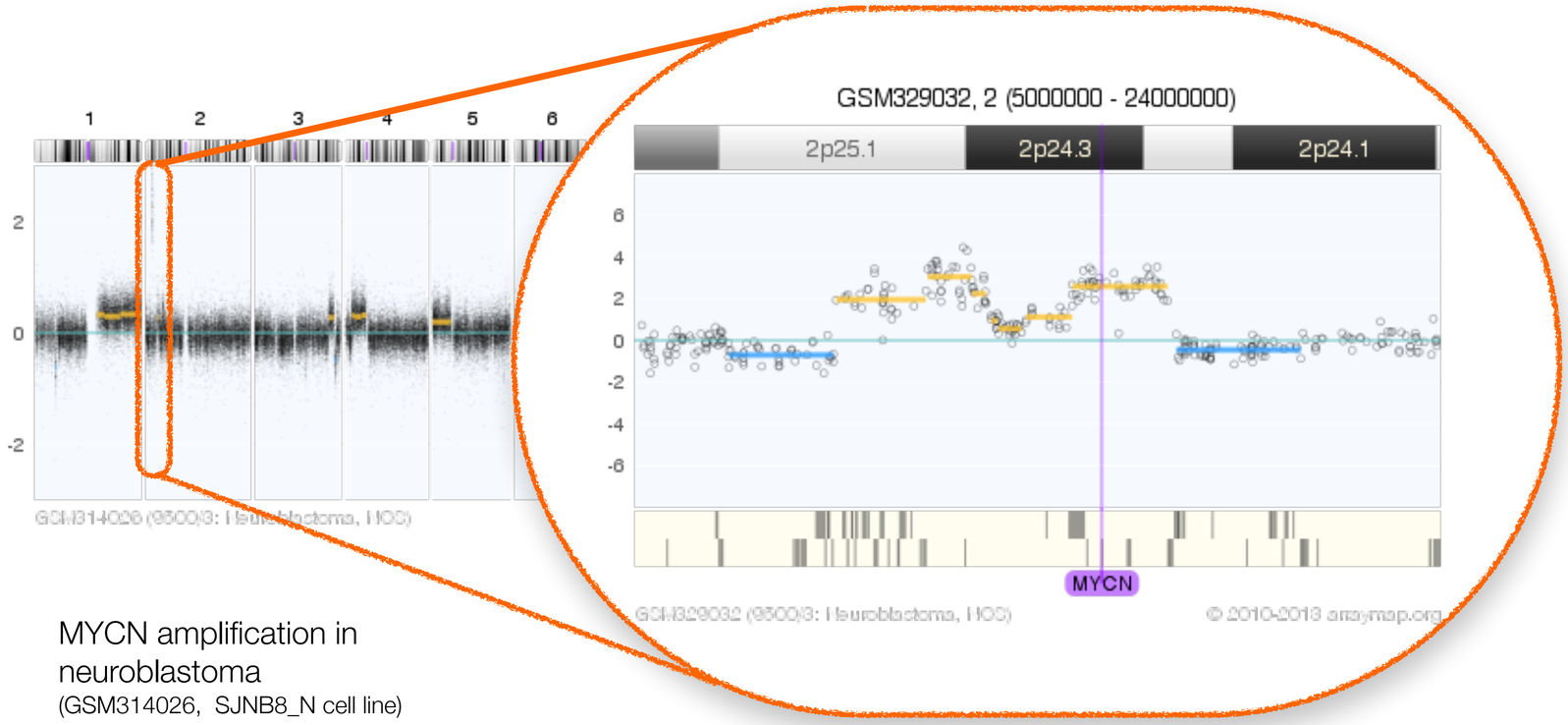
Array-based Detection of Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma



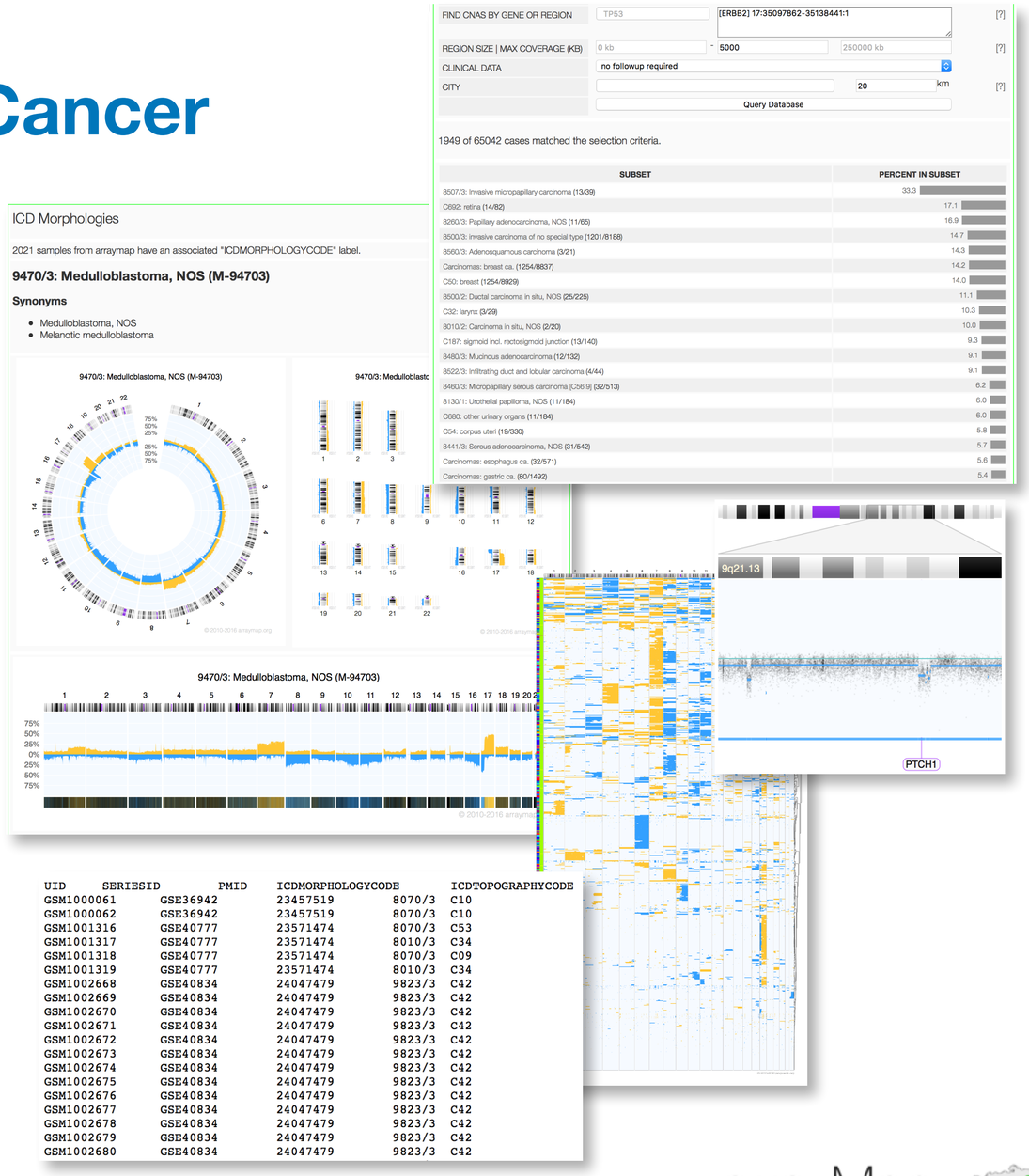
MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

low level/**high level** copy number alterations (CNAs)

arrayMap

arrayMap (2012 - 2020)

Probe-Level Genomic Array Data in Cancer



arrayMap

- Search Samples
- Search Publications
- Progenetix

University of Zurich[UZH]

- Citation & Licensing
- User Guide
- People
- Beacon+

CC BY-SA

Tweet

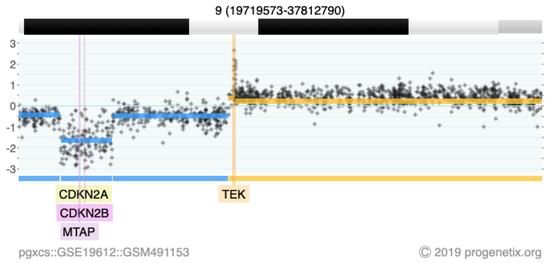
162.158.150.56

visualizing cancer genome array data @ arraymap.org

arrayMap is a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides an entry point for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data.

The current data reflects:

- 72724 genomic array profiles
- 898 experimental series
- 257 array platforms
- ICD-O 341 ICD-O cancer entities
- 795 publications (Pubmed entries)



Genomic copy number imbalances on chromosome 9 in a case of Glioblastoma (**GSM491153**), indcating, among others, a homozygous deletion involving CDKN2A/B.

For the majority of the samples, probe level visualization as well as customized data representation facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools, as we provide through our Progenetix project.

arrayMap is developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich.

RELATED PUBLICATIONS

Cai H, Gupta S, Rath P, Ai N, Baudis M. arrayMap 2014: an updated cancer genome resource. Nucleic Acids Res. 2015 Jan;43(Database issue). Epub 2014 Nov 26.

Cai, H., Kumar, N., & Baudis, M. 2012. arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies. *PLoS One* 7(5), e36944.

Baudis, M. 2007. Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal CGH data. *BMC Cancer* 7:226.

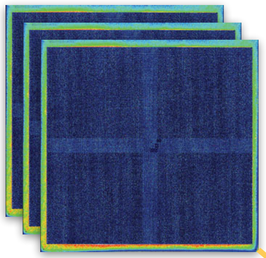
Baudis, M. 2006. Online database and bioinformatics toolbox to support data mining in cancer cytogenetics. *Biotechniques* 40, no. 3: 296-272.

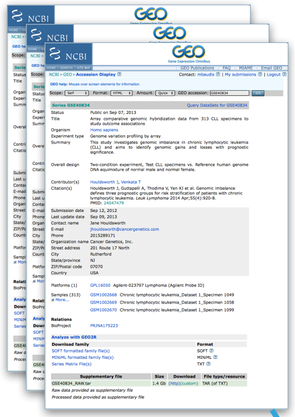
Baudis, M, and ML Cleary. 2001. Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics* 12, no. 17: 1228-1229.

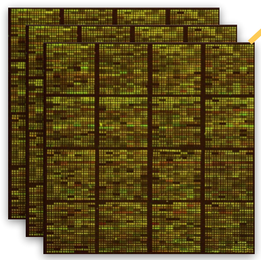
Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project.

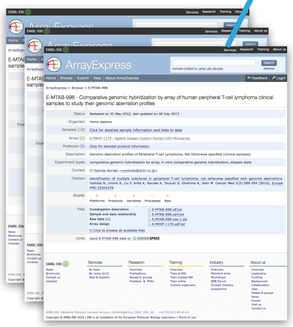
arrayMap

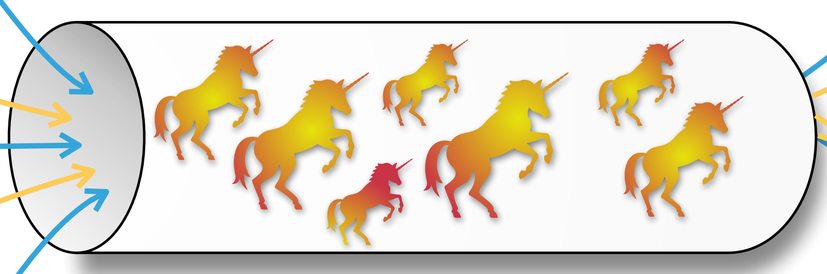
DATA PIPELINE

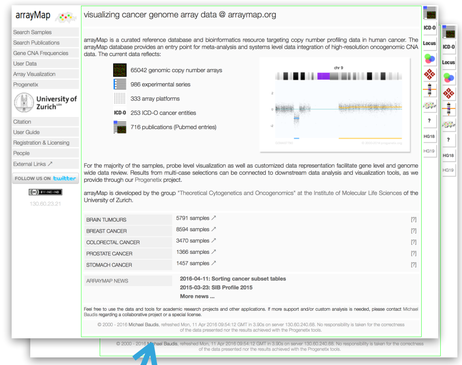


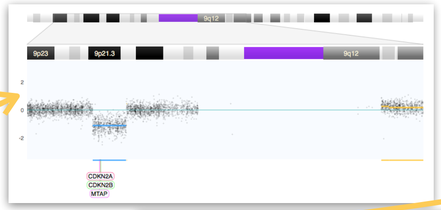


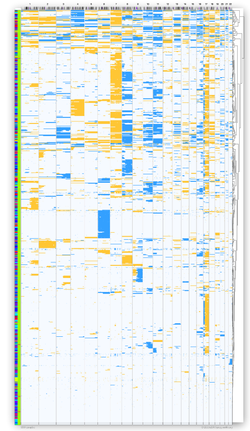


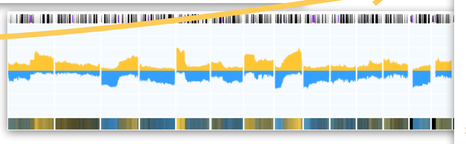


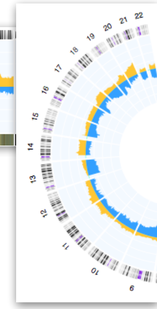


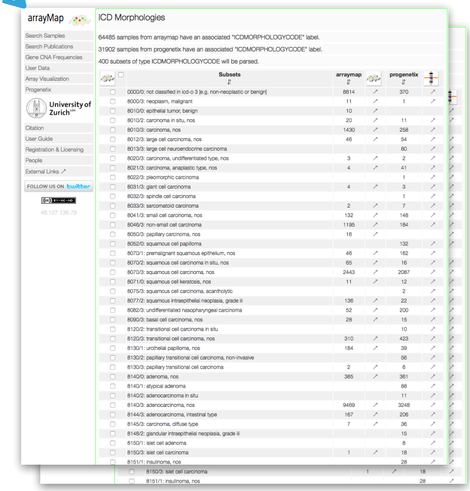












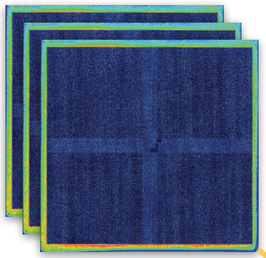
arrayMap

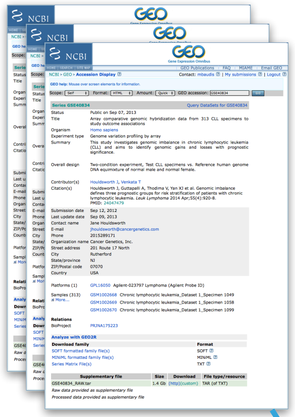
progenetix

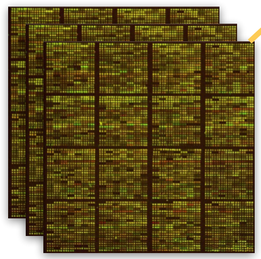
DATA PIPELINE

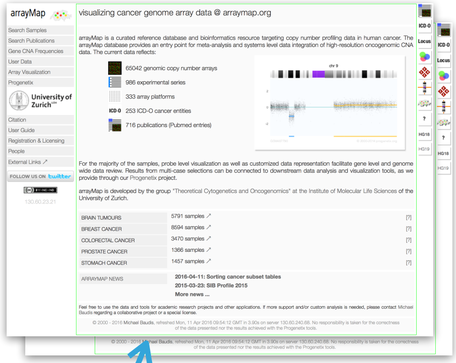
BIOCURATION

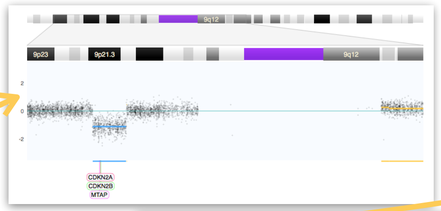
BIOINFORMATICS

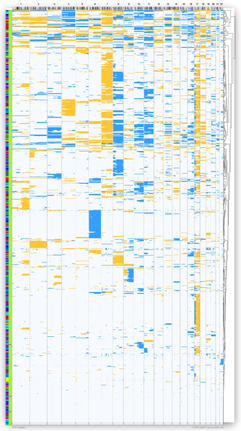


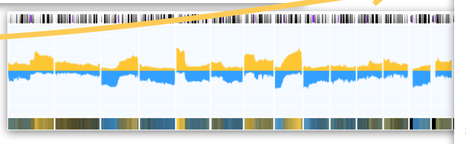


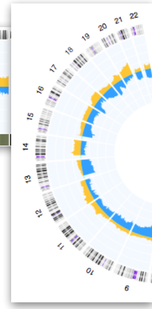


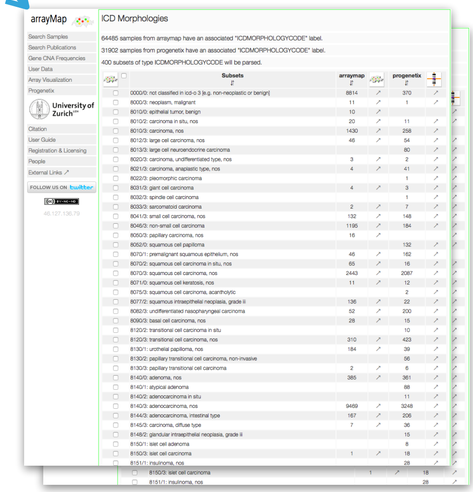








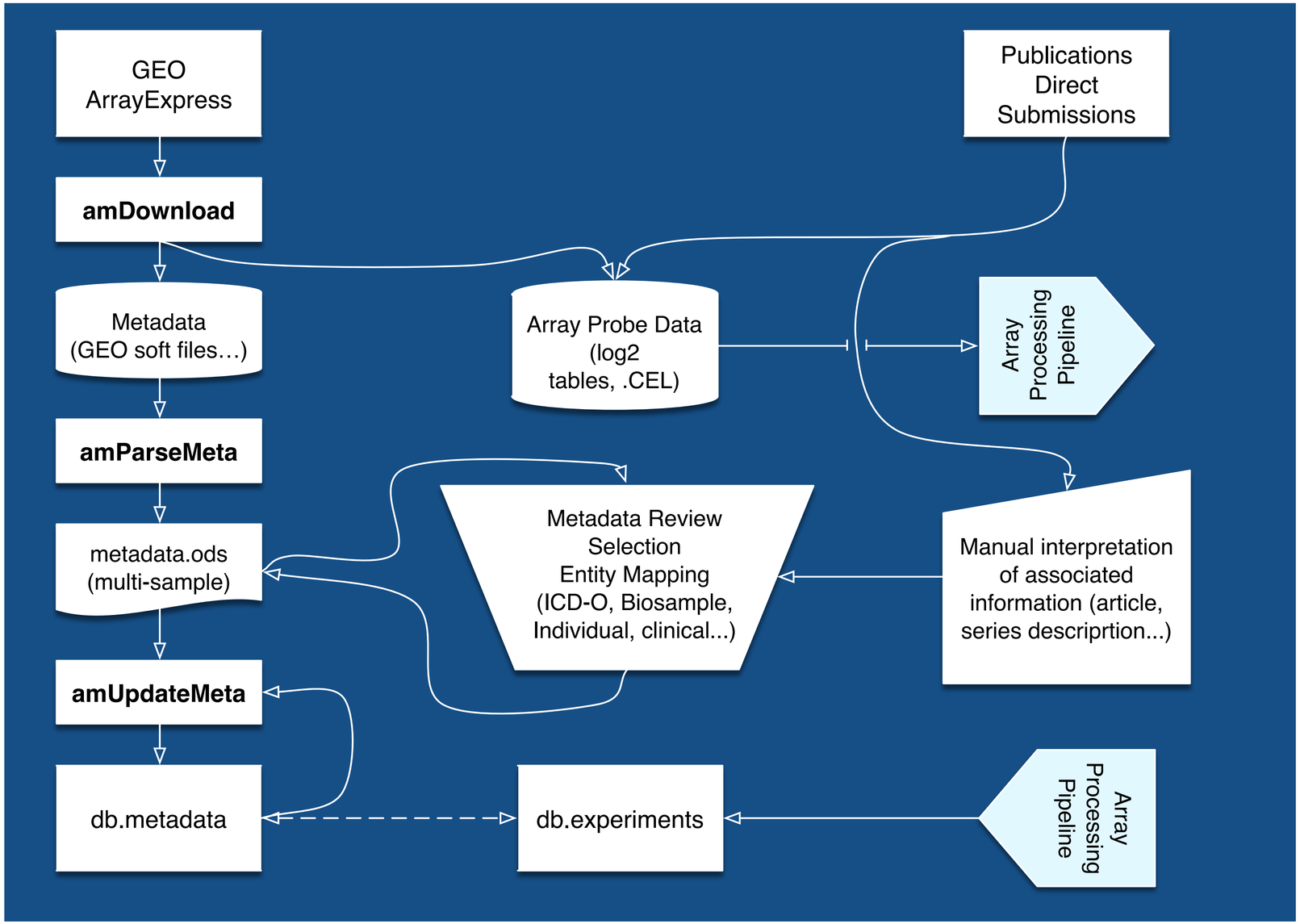




arrayMap

progenetix

Bioinformatics & Data Curation - arrayMap data “Pipeline”



Progenetix & arrayMap: Data Scopes

Biomedical and procedural "Meta"data types

- Diagnostic classification
  - mapping text-based cancer diagnoses to standard classification systems
- Provenance data
  - store identifier-based pointers
  - geographic attribution (individual, biosample, experiment)
- Clinical information
  - **core set** of typical cancer study values:
    - ➡ stage, grade, followup time, survival status, genomic sex, age at diagnosis
  - balance between annotation effort and expected usability

progenetix

Data sets in tutorials

Data sets in the wild

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```
^SAMPLE = GSM174832
!Sample_title = 9194
!Sample_geo_accession = GSM174832
!Sample_status = Public on May 01 2007
!Sample_submission_date = Mar 13 2007
!Sample_last_update_date = Mar 13 2007
!Sample_type = genomic
!Sample_channel_count = 1
!Sample_source_name_ch1 = Bone marrow with 96% blasts
!Sample_organism_ch1 = Homo sapiens
!Sample_taxid_ch1 = 9606
!Sample_characteristics_ch1 = Immunotype: common ALL; Age: 9.2 yrs; Gender: F
!Sample_molecule_ch1 = genomic DNA
!Sample_extract_protocol_ch1 = QiaAmp purification kit (Qiagen)
!Sample_label_ch1 = biotin
!Sample_label_protocol_ch1 = Biotinylated DNA was prepared according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay
manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay manual
701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix hybridisation oven 640 and an Affymetrix Fluidic station 450.
!Sample_scan_protocol = Scanning performed according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3
or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix scanner 3000.
!Sample_description = primary ALL diagnosis sample
!Sample_data_processing = copy number detection using CNAG2.0 software (http://www.genome.umin.jp/)
!Sample_platform_id = GPL3718
!Sample_contact_name = Roland,P.,Kuiper
!Sample_contact_email = r.kuiper@antrg.umcn.nl, e.verwiel@antrg.umcn.nl
!Sample_contact_phone = +31243610868
!Sample_contact_fax = +31243668752
!Sample_contact_department = Human Genetics
!Sample_contact_institute = Radboud University Nijmegen Medical Centre
!Sample_contact_address = Geert Grooteplein 10
!Sample_contact_city = Nijmegen
!Sample_contact_zip/postal_code = 6525GA
!Sample_contact_country = Netherlands
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CEL.gz
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CHP.gz
!Sample_series_id = GSE7255
```

progenetix

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```
^SAMPLE = GSM174832
!Sample_title = 9194
!Sample_geo_accession = GSM174832
!Sample_status = Public on May 01 2007
!Sample_submission_date = Mar 13 2007
!Sample_last_update_date = Mar 13 2007
!Sample_type = genomic
!Sample_channel_count = 1
!Sample_source_name_ch1 = Bone marrow with 96% blasts
!Sample_organism_ch1 = Homo sapiens
!Sample_taxid_ch1 = 9606
!Sample_characteristics_ch1 = Immunotype: common ALL; Age: 9.2 yrs; Gender: F
!Sample_molecule_ch1 = genomic DNA
!Sample_extract_protocol_ch1 = QiaAmp purification kit (Qiagen)
!Sample_label_ch1 = biotin
!Sample_label_protocol_ch1 = Biotinylated DNA was prepared according to the sta
manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard
701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix
!Sample_scan_protocol = Scanning performed according to the standard Affymetrix
or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix scanner 3000
!Sample_description = primary ALL diagnosis sample
!Sample_data_processing = copy number detection using CNAG2.0 software (http://
!Sample_platform_id = GPL3718
!Sample_contact_name = Roland,P.,Kuiper
!Sample_contact_email = r.kuiper@antrg.umcn.nl, e.verwiel@antrg.umcn.nl
!Sample_contact_phone = +31243610868
!Sample_contact_fax = +31243668752
!Sample_contact_department = Human Genetics
!Sample_contact_institute = Radboud University Nijmegen Medical Centre
!Sample_contact_address = Geert Grooteplein 10
!Sample_contact_city = Nijmegen
!Sample_contact_zip/postal_code = 6525GA
!Sample_contact_country = Netherlands
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CEL.gz
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CHP.gz
!Sample_series_id = GSE7255
```

```perl
foreach (grep { ! /characteristics_ch\d/ } @in) {
    my ($key, $value) = split(' = ', $_);
    $key =~ s/[^\w]/_/g;
    if ($key =~ /submission_date/i) {
        $sample->{ YEAR } = $value;
        $sample->{ YEAR } =~ s/^.*?(\d\d\d\d)$/\1/;
    }
```

```perl
$mkey->{ samplekey } = 'AGE';
$mkey->{ matches } = [ qw( age )];

( $mkey->{ retv }, $mkey->{ retk } ) = _grepmeta( $mkey, $meta );

if ($mkey->{ retv } =~ /^(.+?)$/) {
    if ($mkey->{ retv } =~ /month/i) {
        $mkey->{ retk } .= '_months';
        $mkey->{ retv } =~ s/[^\d\.]//g;
    }

    $sample->{ $mkey->{ samplekey } } = _normNumber($mkey->{ retv });
    if ($mkey->{ retk } =~ /month/i) { $sample->{ $mkey->{ samplekey } } /= 12 }
    if ($sample->{ $mkey->{ samplekey } } == 0) { $sample->{ $mkey->{ samplekey } } = 'NA' }
    $sample->{ $mkey->{ samplekey } } = sprintf "%.2f", $sample->{ $mkey->{ samplekey } };
}
```

progenetix

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```
^SAMPLE = GSM286922
!Sample_title = 481 - mAdbID:75320
!Sample_geo_accession = GSM286922
!Sample_status = Public on Sep 04 2008
!Sample_submission_date = May 06 2008
!Sample_last_update_date = Nov 26 2008
!Sample_type = genomic
!Sample_channel_count = 2
!Sample_source_name_ch1 = Normal Lymphocytes
!Sample_organism_ch1 = Homo sapiens
!Sample_taxid_ch1 = 9606
!Sample_characteristics_ch1 = Tissue: lymphocytes
!Sample_molecule_ch1 = genomic DNA
!Sample_extract_protocol_ch1 = Sample DNA Extraction Protocol
!Sample_extract_protocol_ch1 = Other: The DNA was isolated by Qiagen DNe
!Sample_label_ch1 = cy5
!Sample_label_protocol_ch1 = NimbleGen Cy5 Sample Labeling Protocol
!Sample_label_protocol_ch1 = Other: Proprietary protocol information available at http://www.nimblegen.com/technology/index.html
!Sample_source_name_ch2 = 481
!Sample_organism_ch2 = Homo sapiens
!Sample_taxid_ch2 = 9606
!Sample_characteristics_ch2 = Gender: male
!Sample_characteristics_ch2 = Age: 49
!Sample_characteristics_ch2 = Tissue: lymph node
!Sample_characteristics_ch2 = Disease state: Lymphoma
!Sample_characteristics_ch2 = Individual: 481
!Sample_characteristics_ch2 = Clinical info: Submitting diagnosis: DLBCL
!Sample_characteristics_ch2 = Clinical info: Final microarray diagnosis: ABC DLBCL
!Sample_characteristics_ch2 = Clinical info: Follow up status: ALIVE
!Sample_characteristics_ch2 = Clinical info: Follow up years: 10.75
!Sample_characteristics_ch2 = Clinical info: Chemotherapy: CHOP-Like Regimen
!Sample_characteristics_ch2 = Clinical info: ECOG performance status: 2
!Sample_characteristics_ch2 = Clinical info: Stage: 4
!Sample_characteristics_ch2 = Clinical info: LDH ratio: 0.82
!Sample_characteristics_ch2 = Clinical info: Number of extranodal sites: 1
```

Channel 1 is normal -> Cave value swap!

Gender or "chromosomal sex"?

context indicates years, but if it would be a medulloblastoma...

Unknown way to express "alive"!

```
$mkey->{ samplekey } = 'DEATH';
$mkey->{ matches } = [ (
    'death',
    'dead ',
    'vital_status',
    'dead_alive',
    'alive_dead',
) ];

( $mkey->{ retv }, $mkey->{ retk } ) = _grepmeta( $mkey, $meta );

if ($mkey->{ retv } =~ /^(.+?)$/) {
    $sample->{ $mkey->{ samplekey } } = _normDeath($mkey->{ retv }) }
```

Cancer Classifications need an Einstein to sort them out

University of Zurich[UZH]

ARRAYID	NCIT:CODE	NCIT:TERM	ICD-MORPHOLOGY-CODE	ICD-TOPOGRAPHY-CODE
GSM393758	C2773	Acute Myeloid Leukemia Not Otherwise Specified	9861/3	C42
GSM302275	C2852	Adenocarcinoma	8140/3	C34
GSM918983	C3222	Medulloblastoma	9470/3	C7169
GSM551398	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM412374	C3163	Chronic Lymphocytic Leukemia	9823/3	C42
GSM121876	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM714412	C2852	Adenocarcinoma	8140/3	C569
GSM110923	C9306	Soft Tissue Sarcoma	8800/3	C499
GSM711848	C2852	Adenocarcinoma	8140/3	C25
GSM746794	C89476		8077/2	C53
GSM198528	C4017	Ductal Breast Carcinoma	8500/3	C50
GSM271399	C7949		8500/2	C50
GSM533469	C9349	Plasmacytoma	9731/3	C42

Disease annotations in Progenetix

From some text, somewhere, to ontology classes

- **diagnostic categories** are the **most important** labels to associate with genomic observations
- original data almost *never* uses **modern**, **hierarchical** classification systems but provides circumstantial ("breast cancer in pre-menopausal...") or domain-specific ("CLL Binet B", "colorectal carcinoma Dukes C") information
- clinical classifications (ICD-10 ...) have very limited relation to tumor biology
- concepts change over time ...
- for cancer, the "International Classification of Diseases in Oncology" (**ICD-O 3**) by IARC / WHO traditionally has been a good compromise to map to - but with non-hierarchical structure and is used by international reference projects

NCIthesaurus

From Classification to Hierarchical Ontology: ICD-O -> NCIt

example_dx	ICDMORPHOLOGY	ICDOM	ICDTOPOGRAPHY	ICDOT	NCIT:CODE
malignant melanoma [metastatic cell line MaMel19]	Malignant melanoma NOS	8720/3	skin	C44	C3224
malignant melanoma [vagina]	Malignant melanoma NOS	8720/3	vagina and labia	C510	C3224
malignant melanoma [uvea metastasized]	Malignant melanoma NOS	8720/3	retina	C692	C3224
meningioma	Meningioma NOS	9530/0	meninges cerebral spinal	C700	C3230
mesothelioma	Mesothelioma NOS	9050/3	lung and bronchus	C34	C3234
pleural mesothelioma	Mesothelioma NOS	9050/3	pleura	C384	C3234
mesothelioma	Mesothelioma NOS	9050/3	connective and soft tissue NOS	C499	C3234
multiple myeloma	Plasma cell myeloma	9732/3	hematopoietic and reticuloendothelial syste	C42	C3242
Mycosis fungoides	Mycosis fungoides	9700/3	skin	C44	C3246
Myelodysplastic syndrome	Myelodysplastic syndrome NOS	9989/3	hematopoietic and reticuloendothelial syste	C42	C3247
Acute myeloblastic leukemia with maturation [FAB M2]	Acute myeloblastic leukemia with maturation [FAB M2]	9874/3	hematopoietic and reticuloendothelial syste	C42	C3250
neuroblastoma	Neuroblastoma NOS	9500/3	peripheral nervs incl. autonomous	C47	C3270
Cerebral neuroblastoma [cerebral region midline frontal lobe]	Neuroblastoma NOS	9500/3	cerebrum	C710	C3270
neuroblastoma [adrenal gland cell line]	Neuroblastoma NOS	9500/3	adrenal gland	C76	C3270
Cutaneous neurofibroma	Neurofibroma NOS	9540/0	skin	C44	C3272
Plexiform neurofibroma	Neurofibroma NOS	9540/0	Nervous system NOS	C729	C3272
Oligodendroglioma [Supratentorial Frontal Lobe]	Oligodendroglioma NOS	9450/3	cerebrum	C710	C3288
oilgodendroglioma	Oligodendroglioma NOS	9450/3	Brain NOS	C719	C3288
oligodendroglioma	Oligodendroglioma NOS	9450/3	brain nos	c719	C3288
Paraganglioma	Paraganglioma NOS	8680/1	Nervous system NOS	C729	C3308
paraganglioma	paraganglioma NOS	8680/1	adrenal cortex	C740	C3308

- since its beginning Progenetix samples have been classified using the 2 arms of the ICD-O system (morphology ~ histology/biology + topography ~ organ/tissue)
- over the last years we have established mappings between ICD-O code pairs and the NCIt "neoplasm" part of the NCI metathesaurus, thereby empowering hierarchical data structures for search and analysis

DX Ontologies

Hierarchical NCIt Neoplasm Core replaces heterogeneous primary annotations

- heterogeneous and inconsistent diagnostic annotations are common in clinical reports and research studies ("text", ICD-10, ICD-O 3, OncoTree, domain-specific classifications)
- highly **variable granularity** of annotations is a major road block for comparative analyses and large scale data integration
  - "Colorectal Cancer" or "Rectal Mucinous Adenoca."
- initiatives and services such as Phenopackets, MONDO, OXO ... rely on and/or provide mappings to hierarchical ontologies

progenetix

NCIt Neoplasm Core coded display (excerpt) for samples in the Progenetix cancer genome data resource allows sample selection on multiple hierarchy levels →

Collapse all | Expand | All

	Subsets	Samples
☐	NCIT:C3262: Neoplasm	88844
☐	NCIT:C3263: Neoplasm by Site	84747
☐	NCIT:C156482: Genitourinary System Neoplasm	11616
☐	NCIT:C156483: Benign Genitourinary System Neoplasm	219
☐	NCIT:C4893: Benign Urinary System Neoplasm	90
☐	NCIT:C4778: Benign Kidney Neoplasm	90
☐	NCIT:C159209: Kidney Leiomyoma	1
☐	NCIT:C4526: Kidney Oncocytoma	82
☐	NCIT:C8383: Kidney Adenoma	7
☐	NCIT:C7617: Benign Reproductive System Neoplasm	129
☐	NCIT:C4934: Benign Female Reproductive System Neoplasm	129
☐	NCIT:C2895: Benign Ovarian Neoplasm	58
☐	NCIT:C4510: Benign Ovarian Epithelial Tumor	58
☐	NCIT:C40039: Benign Ovarian Mucinous Tumor	58
☐	NCIT:C4512: Ovarian Mucinous Cystadenoma	58
☐	NCIT:C4060: Ovarian Cystadenoma	58
☐	NCIT:C4512: Ovarian Mucinous Cystadenoma	58
☐	NCIT:C3609: Benign Uterine Neoplasm	71
☐	NCIT:C3608: Benign Uterine Corpus Neoplasm	71
☐	NCIT:C3434: Uterine Corpus Leiomyoma	71
☐	NCIT:C156484: Malignant Genitourinary System Neoplasm	11171
☐	NCIT:C157774: Metastatic Malignant Genitourinary System Neoplasm	2
☐	NCIT:C146893: Metastatic Genitourinary System Carcinoma	2
☐	NCIT:C8946: Metastatic Prostate Carcinoma	2
☐	NCIT:C164141: Genitourinary System Carcinoma	10561
☐	NCIT:C146893: Metastatic Genitourinary System Carcinoma	2
☐	NCIT:C8946: Metastatic Prostate Carcinoma	2
☐	NCIT:C3867: Fallopian Tube Carcinoma	19

Standardized Data

Data re-use depends on standardized, machine-readable metadata

- Multiple international initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI ...) work on the generation and implementation of data annotation standards
- emerging / established principles are the use of hierarchical coding systems where individual codes are represented as CURIEs
- other formats for non-categorical annotations based on international standards, e.g.
  - ISO (ISO 8601 time & period, ISO 3166 country codes ...)
  - IETF (GeoJSON ...)
  - W3C (CURIE ...)
- these standards become pervasive throughout GA4GH's ecosystem (e.g. Phenopackets ...)

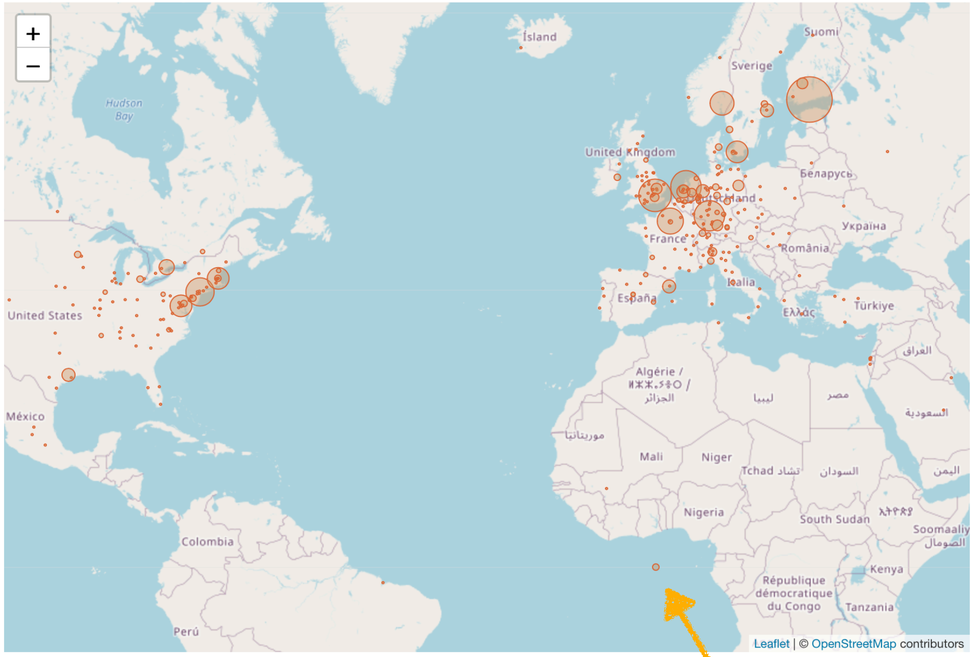
```
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},

"provenance" : {
  "material" : {
    "type" : {
      "id" : "EFO:0009656",
      "label" : "neoplastic sample"
    }
  },
  "geo" : {
    "label" : "Zurich, Switzerland",
    "precision" : "city",
    "city" : "Zurich",
    "country" : "Switzerland",
    "latitude" : 47.37,
    "longitude" : 8.55,
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        8.55,
        47.37
      ]
    },
    "ISO-3166-alpha3" : "CHE"
  }
},
{
  "age": "P25Y3M2D"
}
```

Data Curation

Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

- correct data is important for any type of scientific analysis
- errors in formats and values can occur during all steps between data acquisition and analysis (numerous "Excelgates"!)
- "meta"-resources and analyses are prone to erroneous data due to varying input formats and lack of source control

➡ always look for batch effects and outliers!



Geographic distribution (by corresponding author) of the **118554** genomic array, **36766** chromosomal CGH and **42105** whole genome/exome based cancer genome datasets from the **3306** listed publications. Area sizes correspond to the sample numbers reported from a given location.

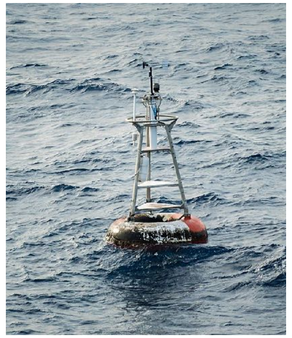
Progenetix publication collection
progenetix.org/publications/list
2020-11-28

25 / 3306 publications

Data Curation - Geolocations

Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

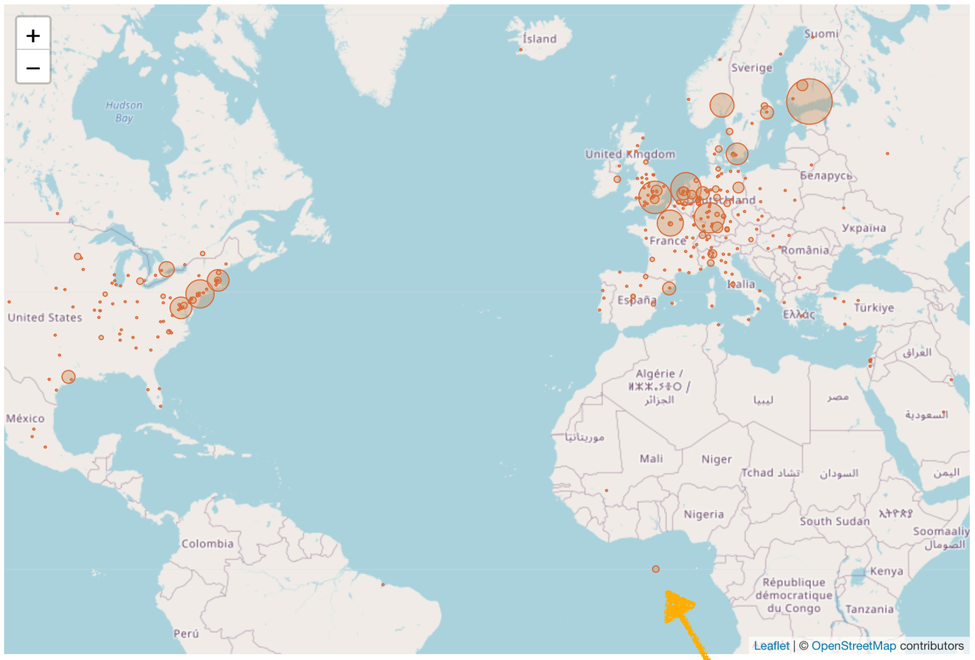
The most geo-tagged place on earth is Null Island



*A troubleshooting country has been added with an Indeterminate sovereignty class called **Null Island (1, 2)**. It is a fictional, 1 meter square island located off Africa where the equator and prime meridian cross. Being centered at 0,0 (zero latitude, zero longitude) it is useful for flagging geocode failures which are routed to 0,0 by most mapping services. Aside: "Null Islands" exist for all local coordinate reference systems besides WGS84 like State Plane (and global if not using modern Greenwich prime meridian). Null Island in Natural Earth is scaleRank 100, indicating it should never be shown in mapping. Side note: Rank 30 (zoom 29 in Google speak)*

https://en.wikipedia.org/wiki/Null_Island

Michael Szell: The Data Science Process 2
http://michael.szell.net/downloads/lecture26_datasciprocess2.pdf
2020-11-25



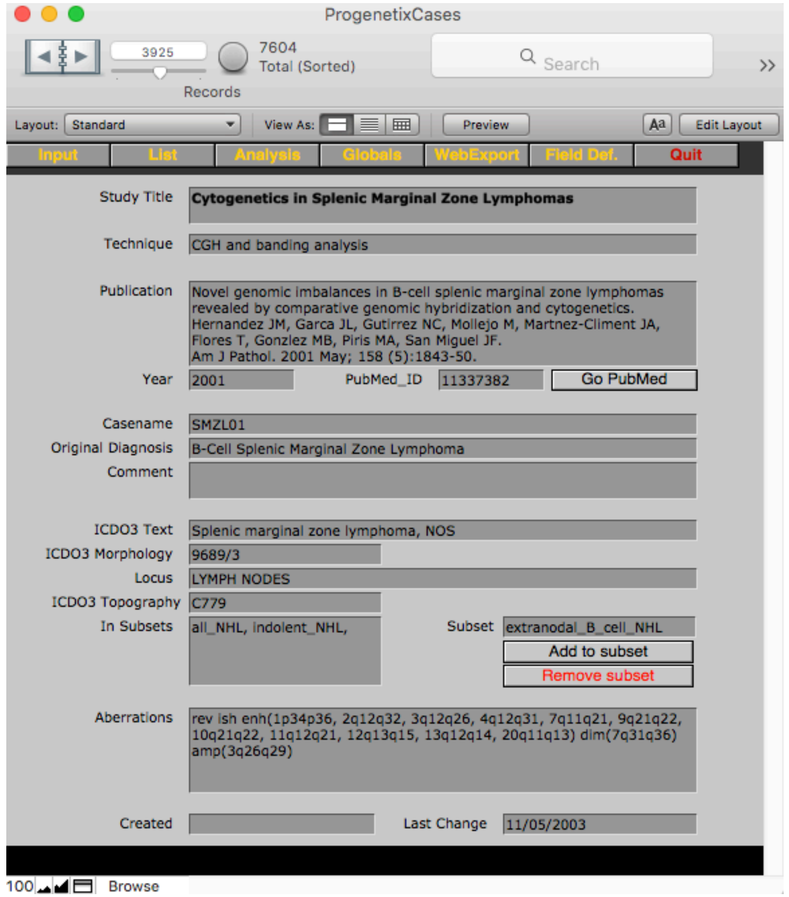
Geographic distribution (by corresponding author) of the **118554** genomic array, **36766** chromosomal CGH and **42105** whole genome/exome based cancer genome datasets from the **3306** listed publications. Area sizes correspond to the sample numbers reported from a given location.

Progenetix publication collection
progenetix.org/publications/list
2020-11-28

25 / 3306 publications

Database Structure

From flat database to hierarchical object storage



Archived version of 2003 "ProgenetixCases" FMP solution

2003

- custom FileMaker database
- text-based annotations
- export & generation of static webpages and data files

2020

- non-SQL document database (MongoDB)
- different object domains connected through identifiers
- data-driven website with JavaScript based frontend and data population through API calls

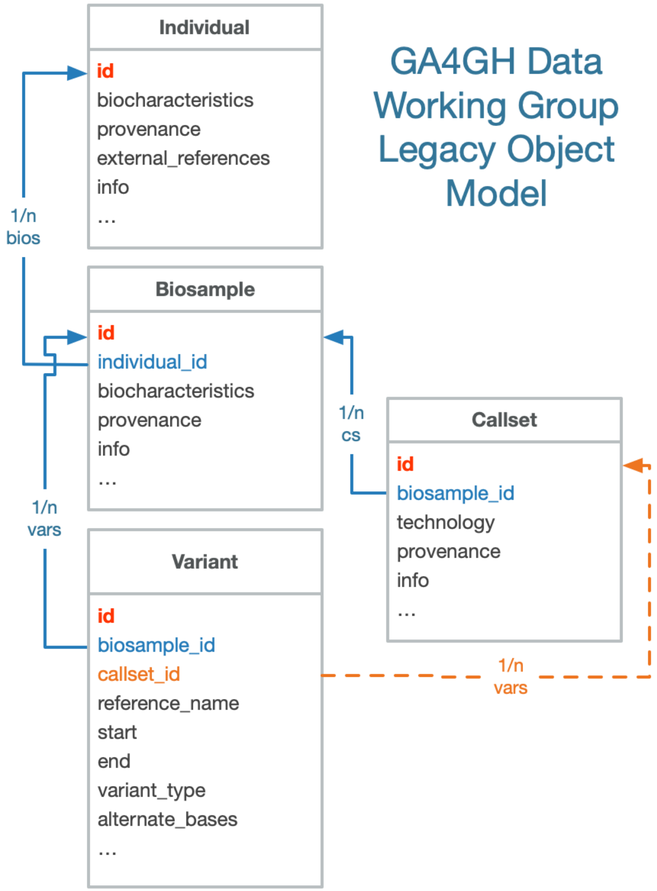

```
{
  "id" : "pgxind-kftx394x",
  "biocharacteristics" : [
    {
      "description" : "female",
      "type" : {
        "id" : "PATO:0020002",
        "label" : "female genotypic sex"
      }
    },
    {
      "description" : null,
      "type" : {
        "id" : "NCBITaxon:9606",
        "label" : "Homo sapiens"
      }
    }
  ],
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  },
  "geo_provenance" : {
    "label" : "Salamanca, Spain",
    "precision" : "city",
    "city" : "Salamanca",
    "country" : "Spain",
    "latitude" : 40.43,
    "longitude" : -3.68
  },
  "info" : {
    "legacy_id" : "PGX_IND_SMZL01"
  },
  "updated" : ISODate("2018-09-26T09:51:39.775Z")
}
```

```
{
  "_id" : ObjectId("5bab583e727983b2e01255ae"),
  "callset_id" : "pgxcs-kftvv618",
  "biosample_id" : "pgxbs-kftvhcao",
  "assembly_id" : "GRCh38",
  "digest" : "7:107200000-158821424:DEL",
  "reference_name" : "7",
  "variant_type" : "DEL",
  "start" : 107200000,
  "end" : 158821424,
  "info" : {
    "cnv_value" : null,
    "cnv_length" : 51621424
  },
  "updated" : "2018-09-26 09:51:39.775397"
}
```

```
{
  "_id" : ObjectId("5bab56cd727983b2e00b0bde"),
  "id" : "pgxbs-kftvhcao",
  "description" : "Splenic Marginal Zone Lymphoma",
  "biocharacteristics" : [
    {
      "type" : {
        "id" : "UBERON:0002106",
        "label" : "spleen"
      }
    },
    {
      "type" : {
        "id" : "icdot-C42.2",
        "label" : "Spleen"
      }
    },
    {
      "type" : {
        "id" : "icdom-96893",
        "label" : "Splenic marginal zone B-cell lymphoma"
      }
    },
    {
      "type" : {
        "id" : "NCIT:C4663",
        "label" : "Splenic Marginal Zone Lymphoma"
      }
    }
  ],
  "individual_id" : "pgxind-kftx394x",
  "individual_age_at_collection" : "P67Y",
  "info" : {
    "death" : "0",
    "followup_months" : 53,
    "callset_ids" : [
      "pgxcs-kftvv618"
    ],
    "legacy_id" : "PGX_AM_BS_SMZL01"
  },
  "external_references" : [
    {
      "type" : {
        "id" : "PMID:11337382"
      }
    }
  ],
  "provenance" : {
    "material" : {
      "type" : {
        "id" : "EFO:0009656",
        "label" : "neoplastic sample"
      }
    },
    "geo" : {
      "label" : "Salamanca, Spain",
      "precision" : "city",
      "city" : "Salamanca",
      "country" : "Spain",
      "geojson" : {
        "type" : "Point",
        "coordinates" : [
          -3.68,
          40.43
        ]
      },
      "ISO-3166-alpha3" : "ESP"
    }
  },
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  }
}
```

Database Structure

From flat database to hierarchical object storage



GA4GH Data Working Group Legacy Object Model

- collections in Progenetix MongoDB database reflect a consensus domain model for genomic data repositories
- flexible linking and object structure facilitates rapid change-overs
- BSON/JSON format in DB
  - equals data in JavaScript
  - "equals" objects in Python, Perl

➡ **rapid prototyping and implementation**

2020

- non-SQL document database (MongoDB)
- different object domains connected through identifiers
- data-driven website with JavaScript based frontend and data population through API calls


```
{
  "id" : "pgxind-kftx394x",
  "biocharacteristics" : [
    {
      "description" : "female",
      "type" : {
        "id" : "PATO:0020002",
        "label" : "female genotypic sex"
      }
    },
    {
      "description" : null,
      "type" : {
        "id" : "NCBITaxon:9606",
        "label" : "Homo sapiens"
      }
    }
  ],
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  },
  "geo_provenance" : {
    "label" : "Salamanca, Spain",
    "precision" : "city",
    "city" : "Salamanca",
    "country" : "Spain",
    "latitude" : 40.43,
    "longitude" : -3.68
  },
  "info" : {
    "legacy_id" : "PGX_IND_SMZL01"
  },
  "updated" : ISODate("2018-09-26T09:51:39.775Z")
}
```

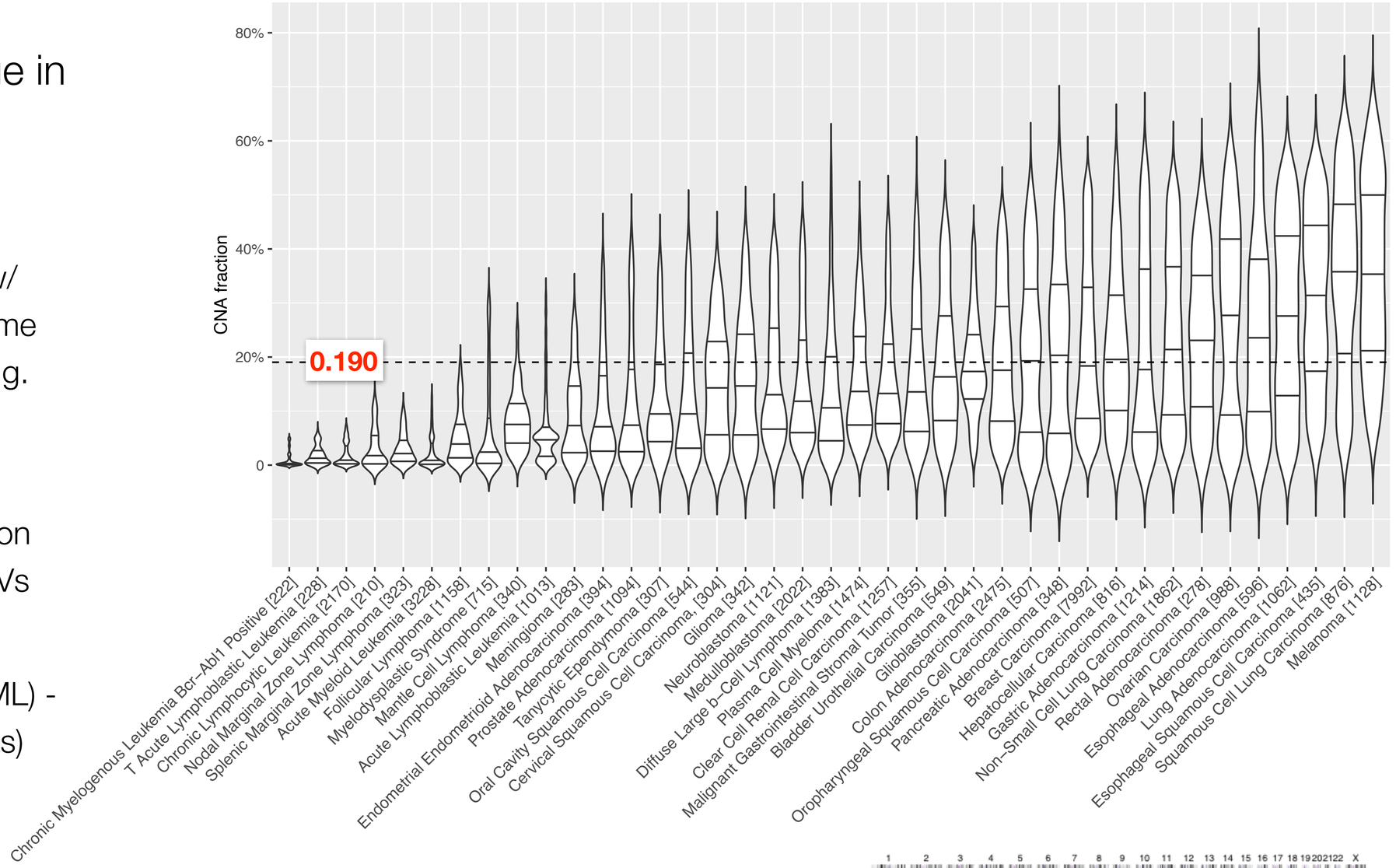
```
{
  "_id" : ObjectId("5bab583e727983b2e01255ae"),
  "callset_id" : "pgxcs-kftvv618",
  "biosample_id" : "pgxbs-kftvhcao",
  "assembly_id" : "GRCh38",
  "digest" : "7:107200000-158821424:DEL",
  "reference_name" : "7",
  "variant_type" : "DEL",
  "start" : 107200000,
  "end" : 158821424,
  "info" : {
    "cnv_value" : null,
    "cnv_length" : 51621424
  },
  "updated" : "2018-09-26 09:51:39.775397"
}
```

```
{
  "_id" : ObjectId("5bab56cd727983b2e00b0bde"),
  "id" : "pgxbs-kftvhcao",
  "description" : "Splenic Marginal Zone Lymphoma",
  "biocharacteristics" : [
    {
      "type" : {
        "id" : "UBERON:0002106",
        "label" : "spleen"
      }
    },
    {
      "type" : {
        "id" : "icdot-C42.2",
        "label" : "Spleen"
      }
    },
    {
      "type" : {
        "id" : "icdom-96893",
        "label" : "Splenic marginal zone B-cell lymphoma"
      }
    },
    {
      "type" : {
        "id" : "NCIT:C4663",
        "label" : "Splenic Marginal Zone Lymphoma"
      }
    }
  ],
  "individual_id" : "pgxind-kftx394x",
  "individual_age_at_collection" : "P67Y",
  "info" : {
    "death" : "0",
    "followup_months" : 53,
   "callset_ids" : [
      "pgxcs-kftvv618"
    ],
    "legacy_id" : "PGX_AM_BS_SMZL01"
  },
  "external_references" : [
    {
      "type" : {
        "id" : "PMID:11337382"
      }
    }
  ],
  "provenance" : {
    "material" : {
      "type" : {
        "id" : "EFO:0009656",
        "label" : "neoplastic sample"
      }
    },
    "geo" : {
      "label" : "Salamanca, Spain",
      "precision" : "city",
      "city" : "Salamanca",
      "country" : "Spain",
      "geojson" : {
        "type" : "Point",
        "coordinates" : [
          -3.68,
          40.43
        ]
      },
      "ISO-3166-alpha3" : "ESP"
    }
  },
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  }
}
```

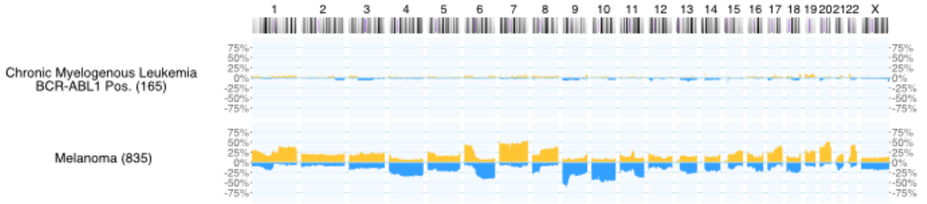
Data Use Cases

Genome CNV coverage in Cancer Classes

- 43654 out of 93640 CNV profiles; filtered for entities w/ >200 samples (removed some entities w/ high CNV rate, e.g. sarcoma subtypes)
- Single-sample CNV profiles were assessed for the fraction of the genome showing CNVs (relative gains, losses)
- range of medians 0.001 (CML) - 0.358 (malignant melanomas)



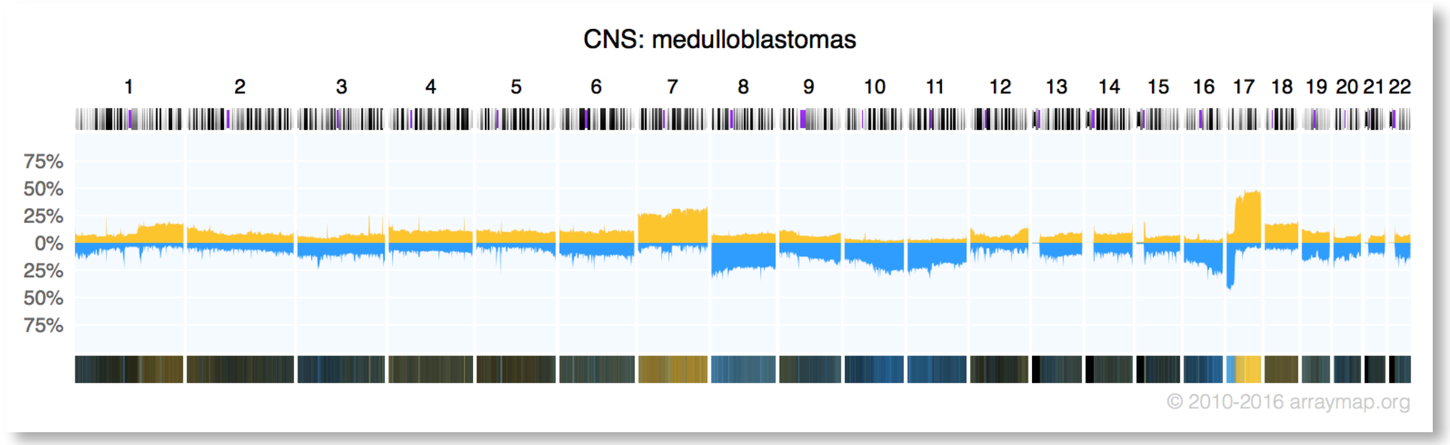
Lowest / Highest CNV fractions =>

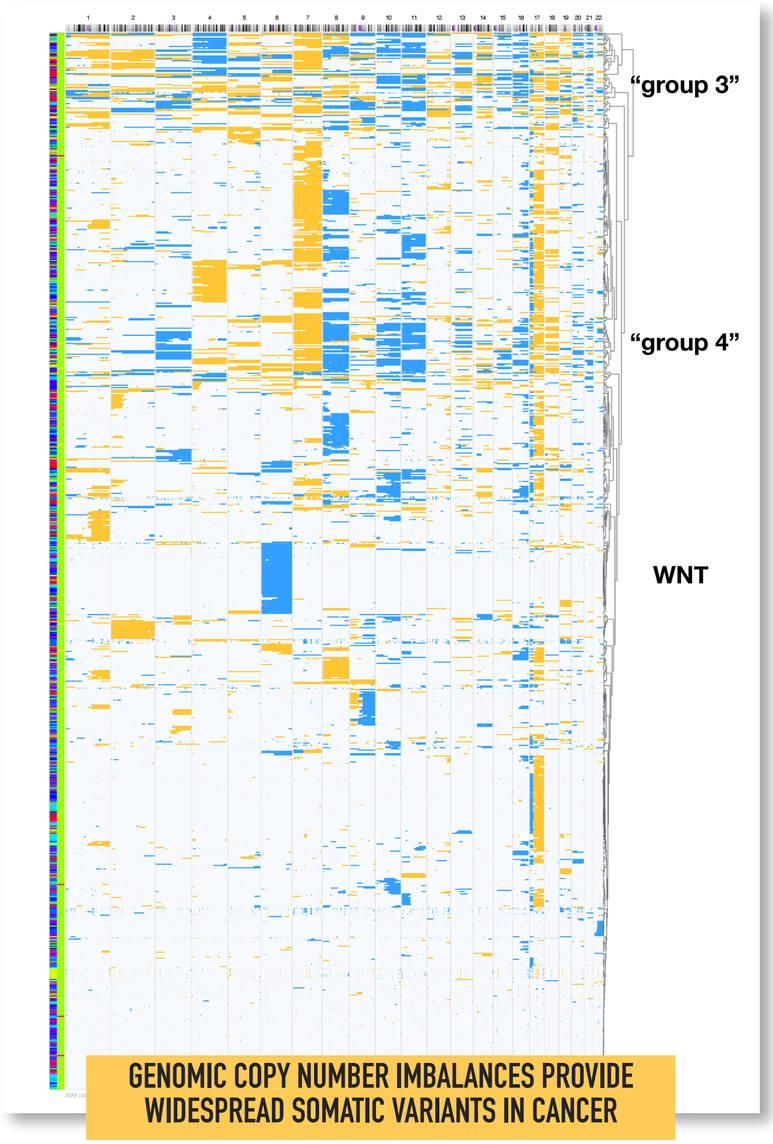


Somatic CNVs In Cancer

Recurrent mutation patterns

How can those patterns be used for classification and determination of biological mechanisms?

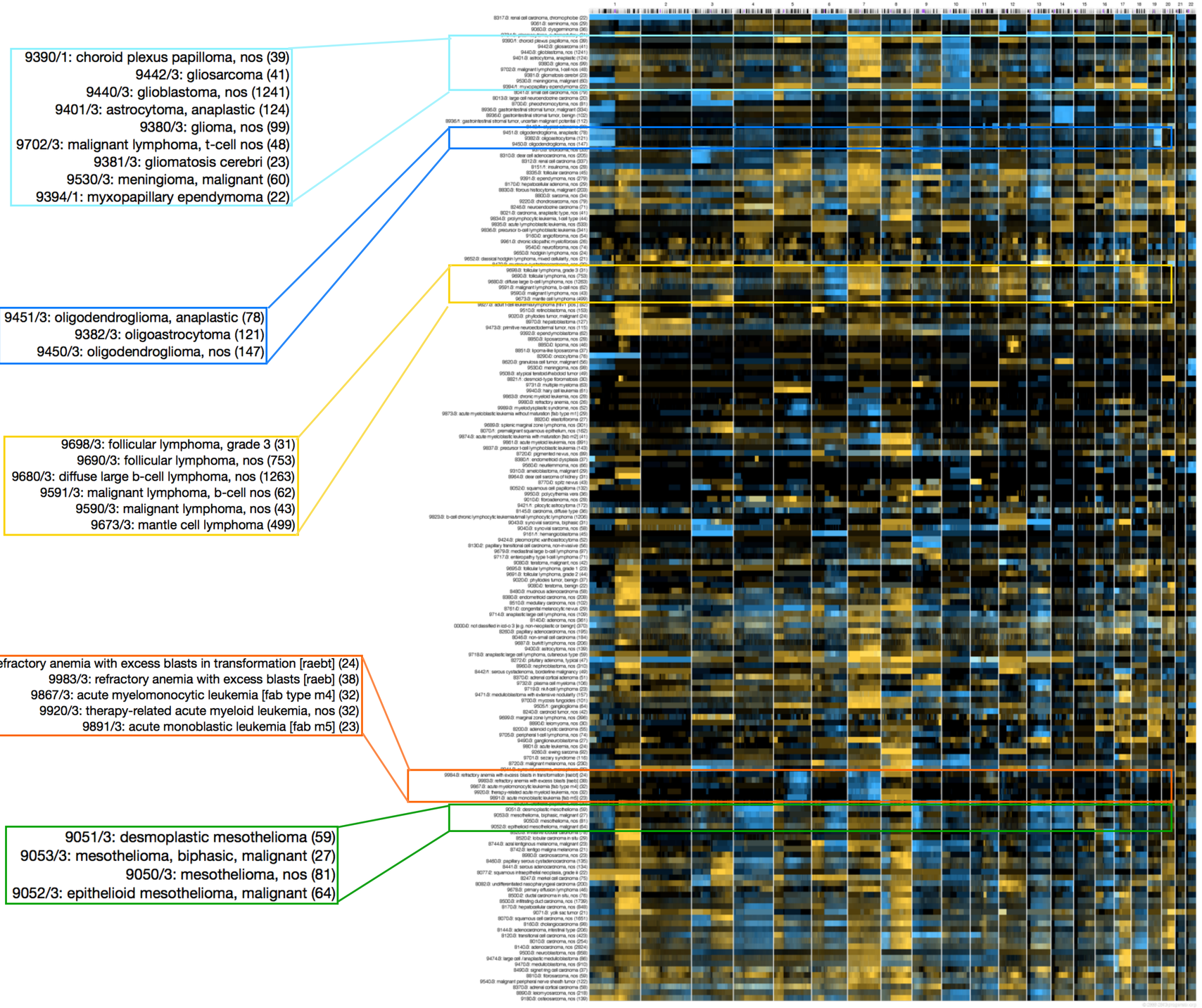




A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional relevance and clinical correlation.

arrayMap

Somatic Mutations In Cancer: Patterns
Making the case for genomic classifications
Some related cancer entities show similar copy number profiles



arrayMap

BIOINFORMATICS ORIGINAL PAPER

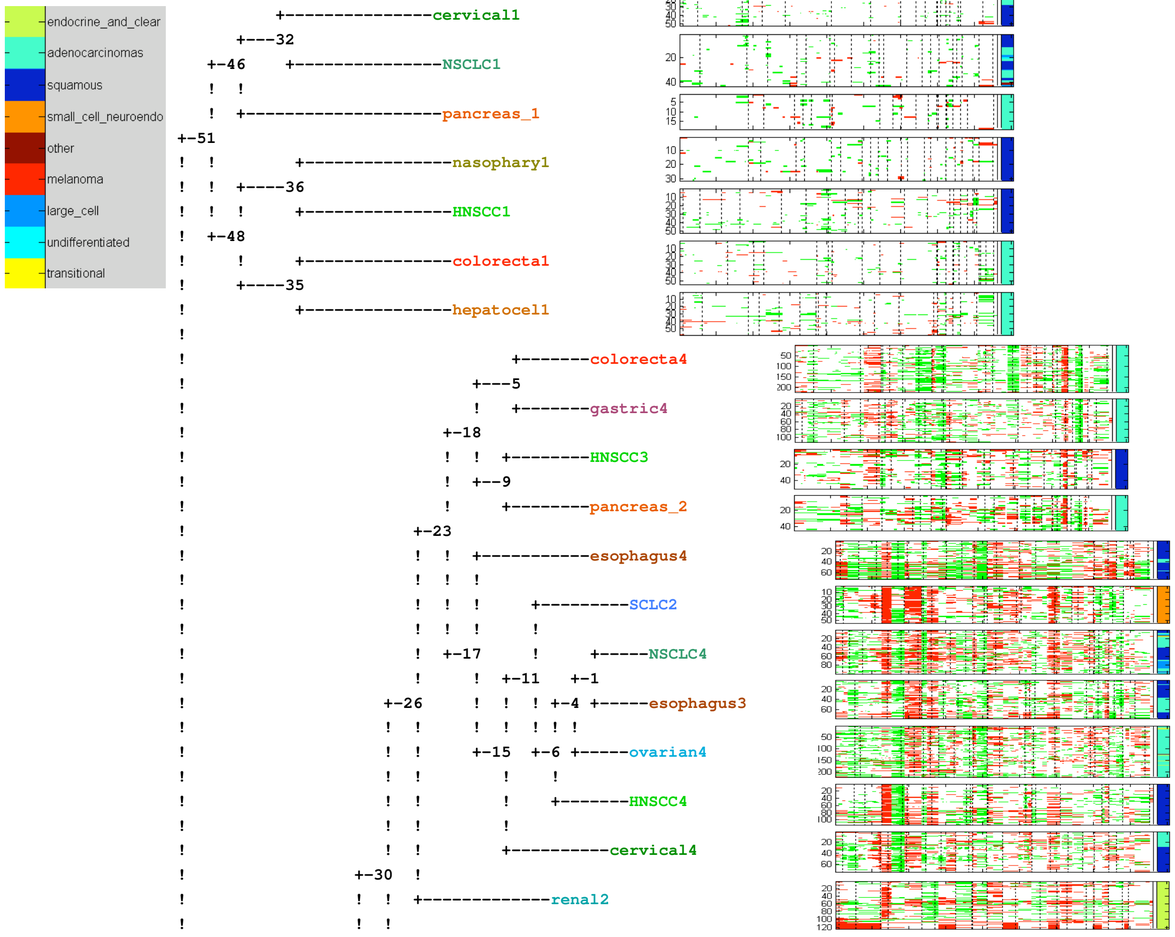
Vol. 25 no. 17 2009, pages 2208–2215
doi:10.1093/bioinformatics/btp365

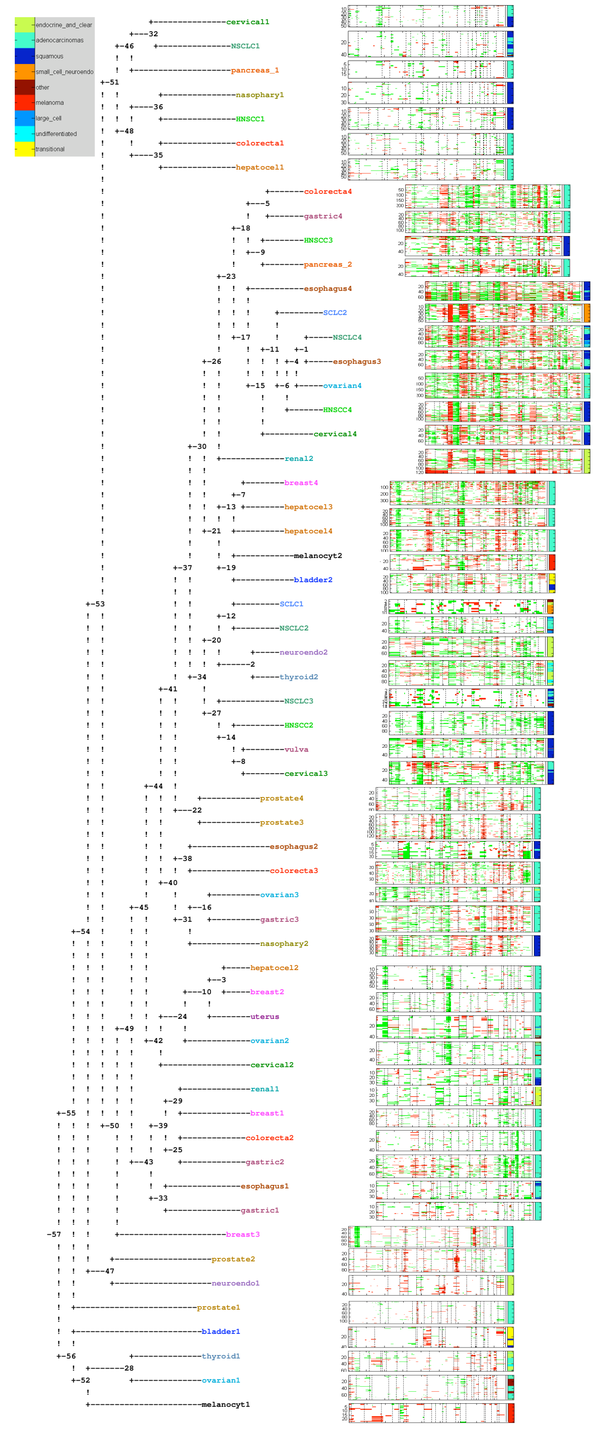
Gene expression

Inferring progression models for CGH data

Jun Liu[1], Nirmalya Bandyopadhyay[1,*], Sanjay Ranka[1], M. Baudis[2] and Tamer Kahveci[1,*]

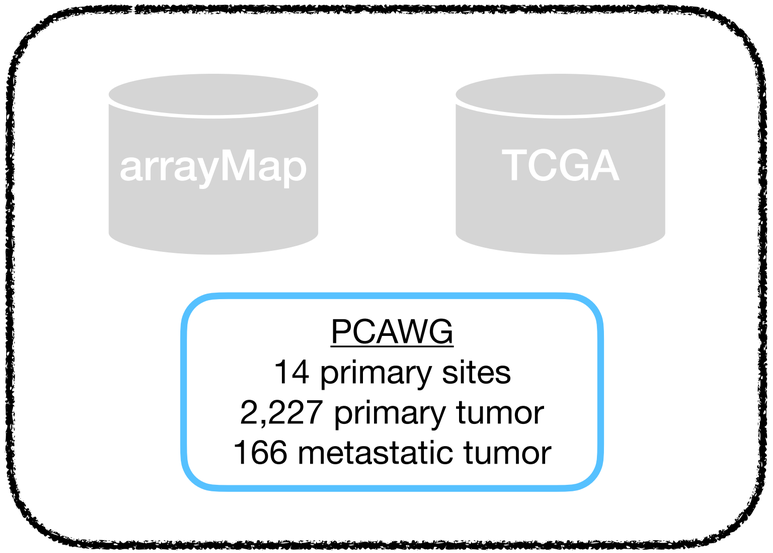
[1]Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA and [2]Institute for Molecular Biology, University of Zurich, Zurich, Switzerland





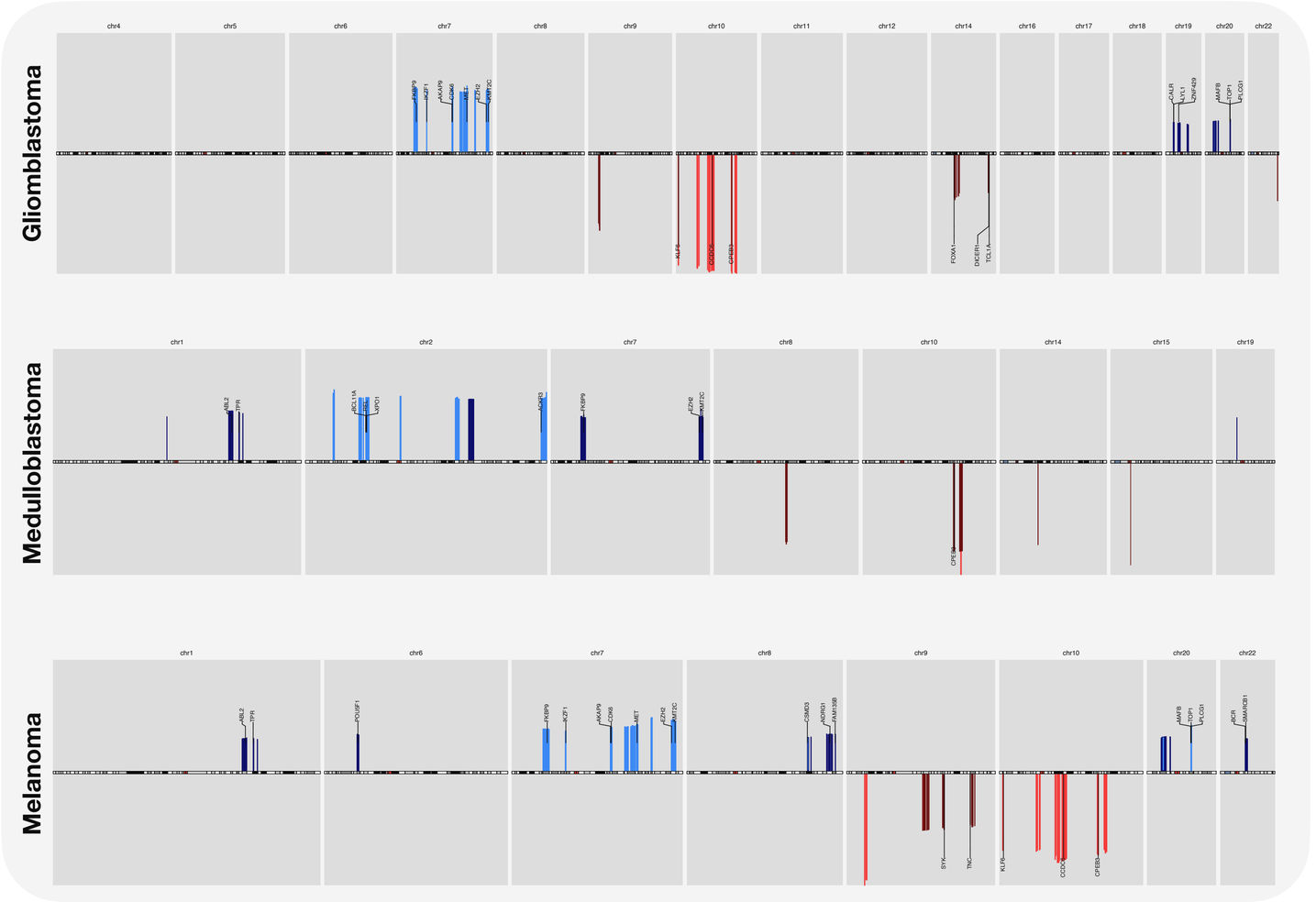
Unique Patterns of Copy Number Mutations Across Cancer Types

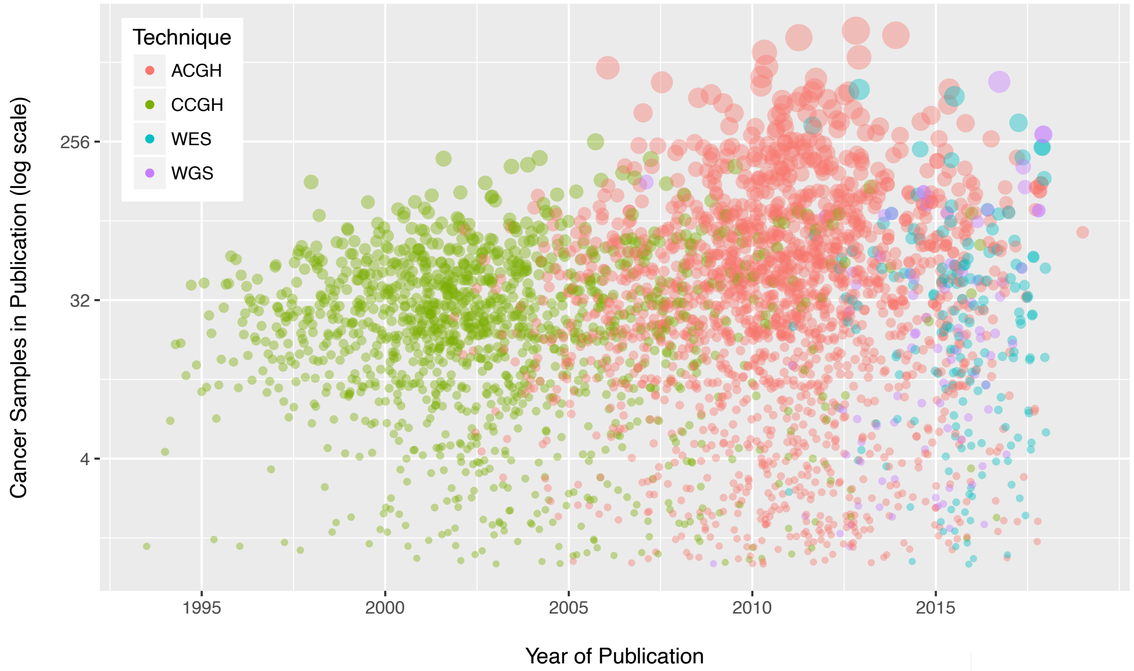
An extensive collection of tumor CNV



- **Unique & distinctive patterns**
- **Patterns of sites and disease**
- **Identify the origin of a tumor**

Examples of unique CNV patterns





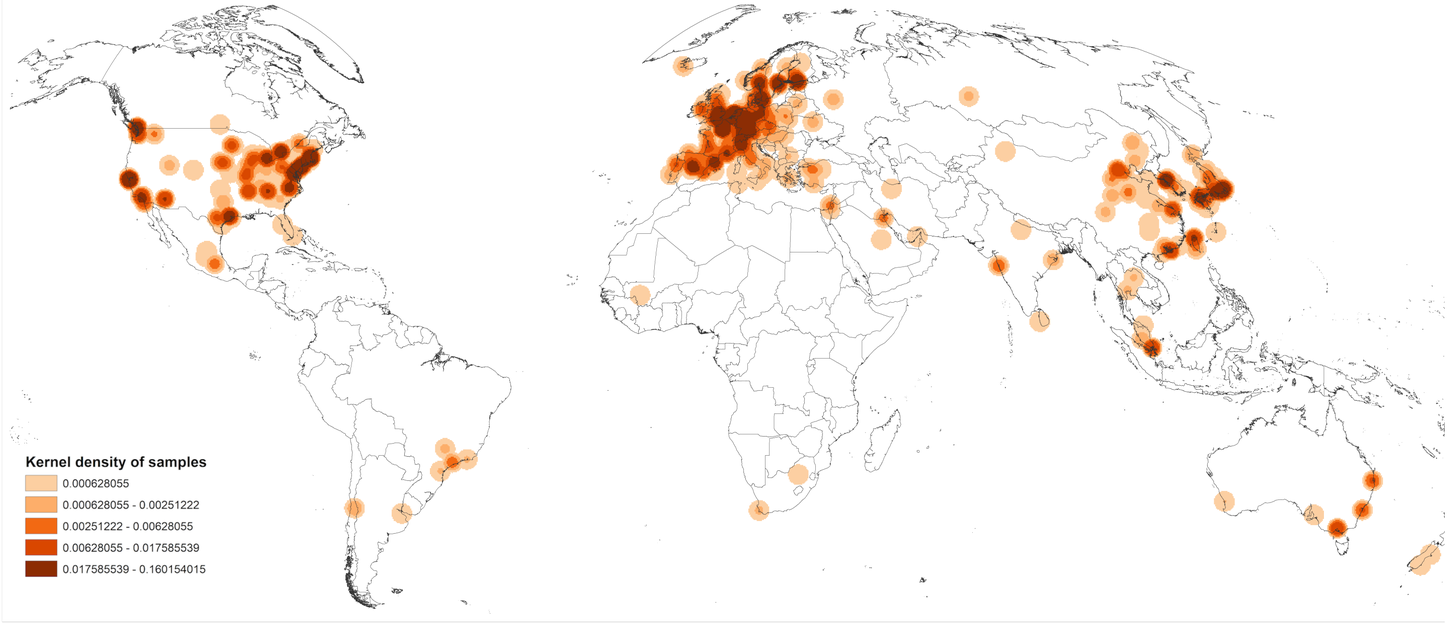
Publication Landscape of Cancer CNV Profiling

Publication statistics for cancer genome screening studies. The graphic shows our as- sessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

Map of the geographic distribution (by first author affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets.

The numbers are derived from the 3'240 publications registered in the Progenetix database.



progenetix

Progenetix as Example Genomics Resource

Some trajectories ...

- from local database to **online resource**
- from flat database to **hierarchical object storage**
- from dedicated database to mix of **open software tools**
- from static pages to **data driven website**
- from copy, paste, clean to **automated download & process** - still edit & clean
- from registered access to raw data & commercial licensing to **CC BY 4.0** (CC0 for tools)
- from local software development to **open code on Github**
- from standalone resource to federated data, **APIs** and services

progenetix

(Bio)informatics Skill Set

What has been needed to develop & maintain progenetix.org?

- Scripting and application development using Python, Perl and JavaScript
- Data analysis and plotting in R, Python and Perl
- Regular expressions for data entry an (programmatic) identifier matching
- JSON, YAML, tab-delimited text as file formats; some binary source files (.CEL)
- non-SQL database (MongoDB) for flexibility and document structure
- web development with Perl, Python, JS, React and Apache server; Cloudflare
- No proprietary software involved (some OpenOffice Calc / Google Sheets spreadsheets for data cleanup)

progenetix

(Bio)informatics Skill Set

What has been needed to develop & maintain progenetix.org?

GitHub

text mining

Perl

regular expressions

```
s/knowledge/mastery/
```

React

mongoDB

APACHE HTTP SERVER PROJECT

R

array pipelines

statistics

python™

HTML CSS JS

Bachelor / Master Project in Data Wrangling? Ask!

BIO390 HS20

Exam planning

- On site exam!
- 2020-12-15
- different lecture hall: **Y-24G-45**
- time: 08:15-09:15
  - limiting on-site exposure time (60 instead of 90 minutes)
- multiple (single + multiple) choice w/ one or two open questions
- no material, phones etc.
- student ID for entrance