Building Bioinformatics Resources

Make quantitative biological data accessible

Qingyao Huang | UZH BIO390 HS24

Learning objectives

- What comprises a bioinformatics resource?
- What types of data are typically included?
- How to use the resources?
- How to build / maintain a resource?
- What tools / skills are needed?

Primary data repositories

- Collects primary datasets conducted by individual researchers
- Often required by publication, as part of reproducibility, open data effort
- Examples:
  - Gene Expression Omnibus (GEO; NCBI)
  - ArrayExpress (EBI)
  - GenBank
  - Protein Data Bank
  - Proteomics Identification Database (PRIDE; EBI)

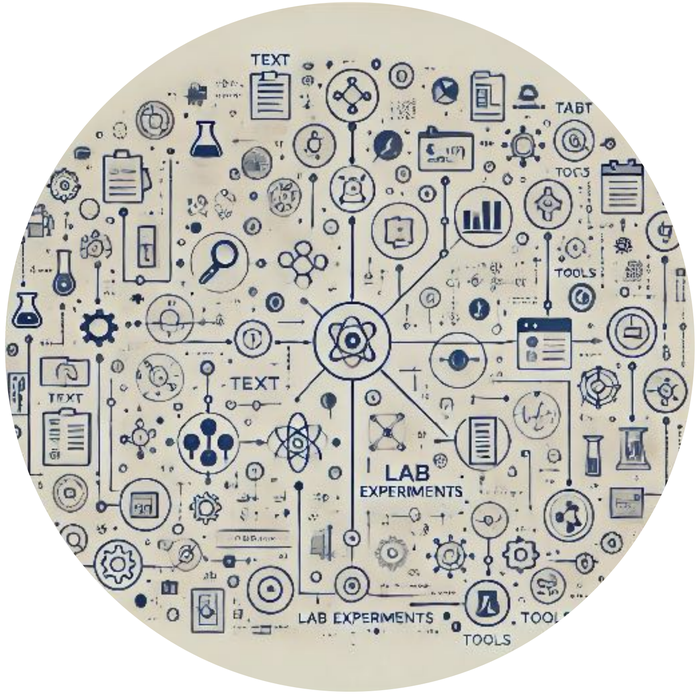
Curated databases

- Codify terms, classifications, based on existing knowledge derived from primary research
- Examples:
  - Kyoto Encyclopedia of Genes and Genomes (KEGG)
  - Gene Ontology (GO)
  - Reactome
  - Most ontologies, e.g.
    - Disease ontology
    - BRENDA tissue ontology
    - Cell Line Ontology …

Bioinformatics online tools

- Web service often with computation support, so the user can access the information through browsing or upload the data for analysis without necessarily possessing programming skills or set up computation locally
- Examples:
  - UCSC genome browser
  - NCBI BLAST (Basic Local Alignment Search Tool)
  - Galaxy (workflow) server
  - Enrichment services

Meta-databases / knowledge-bases



- Integrates many primary datasets and their metadata (type of study, experimental set-up/ replicates, sample conditions…) with textmining techniques
- Examples:
  - Progenetix
  - PaxDb
  - STRING
  - UniProtKB

Progenetix

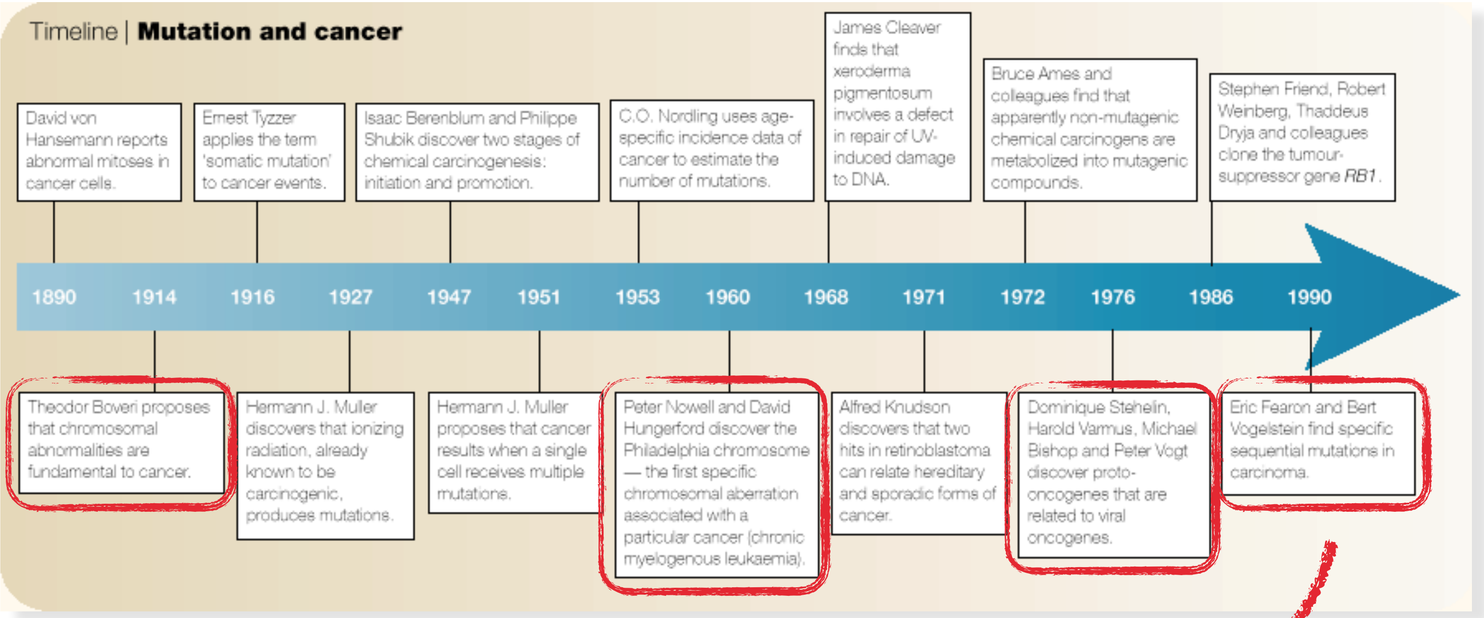
- Motivation for building a resource
  - History
  - Relevance
- What is the quantitative data and how is it represented?
  - Copy number variation
  - Techniques
- What is the metadata?
  - Codify cancer types, stage, patient information
  - Geographical information
- How to access the data safely?
  - Sensitive human data

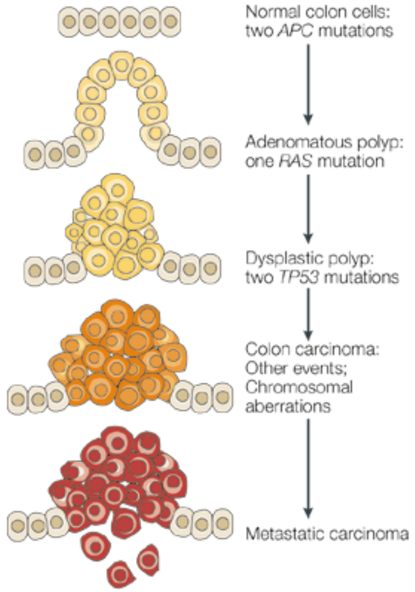
Building a Genomics Resource

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



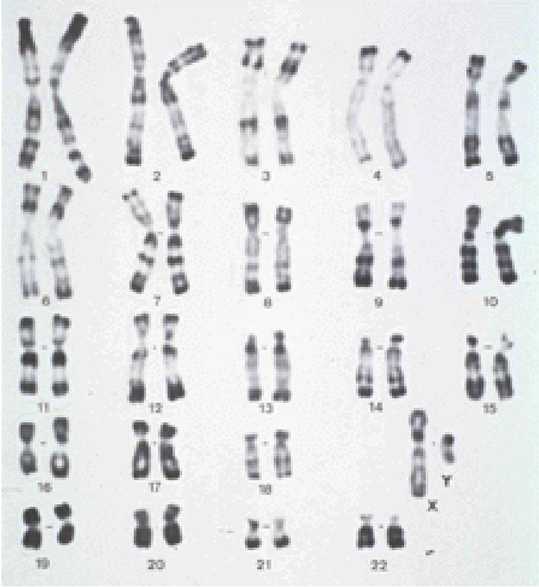


Cancers are based on acquired and inherited genomic mutations

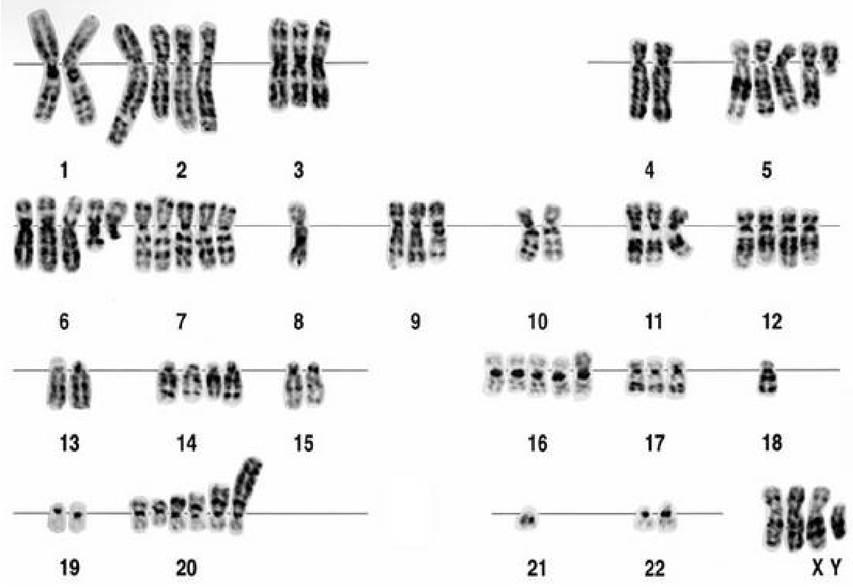
Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. Nature Reviews Cancer, 1(2), 157–162.

Genomic changes at the DNA level are hallmarks of cancer

We inherited 23 paternal and 23 maternal chromosomes, mostly identical.



Normal karyotype



Tumor karyotype

Our goal: identify CN changes to improve characterization, classification, and treatment of cancers

Janet Rowley (1970s)

Chromosomal translocations in cancer

- Recurrent chromosomal translocations in leukemias / lymphomas
- "**Philadelphia chromosome**" in CML (Nowell & Hungerford, 1960) abnormally short chromosome 22
- 1972: t(8;21) ALL: AML1-ETO fusion protein
- 1973: t(9;22) CML: BCR-ABL fusion protein
- 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)

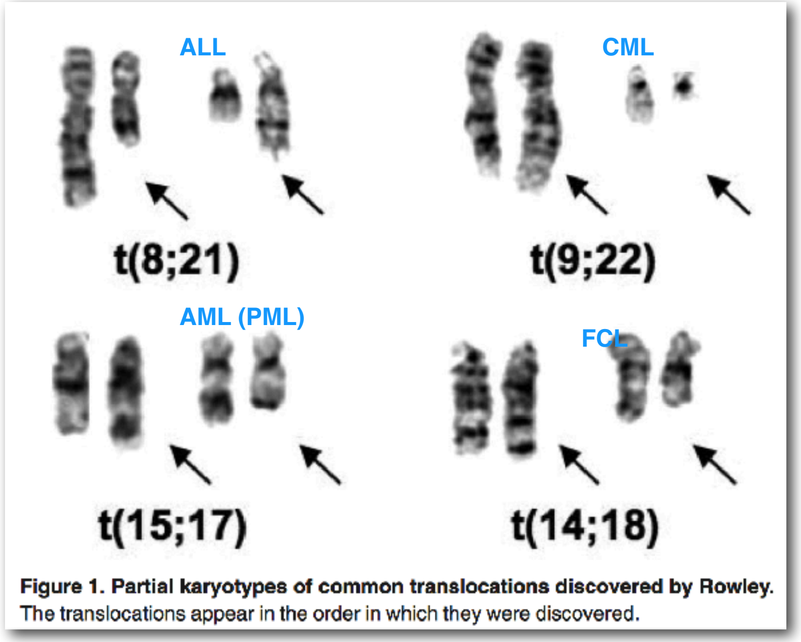
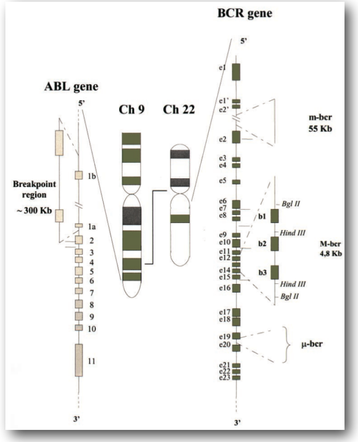
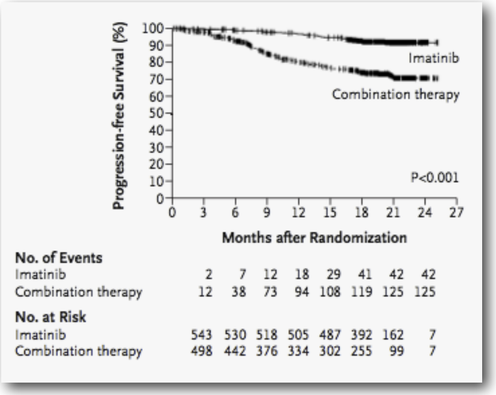


Figure 1. Partial karyotypes of common translocations discovered by Rowley. The translocations appear in the order in which they were discovered.

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)

University of Zurich[UZH]

Department of Molecular Life Sciences

Genome screening at the core of “Personalised Health”

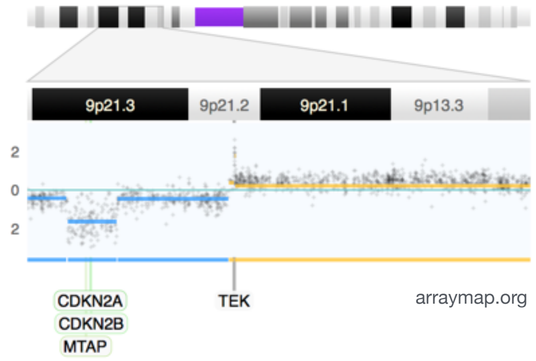
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**
- **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

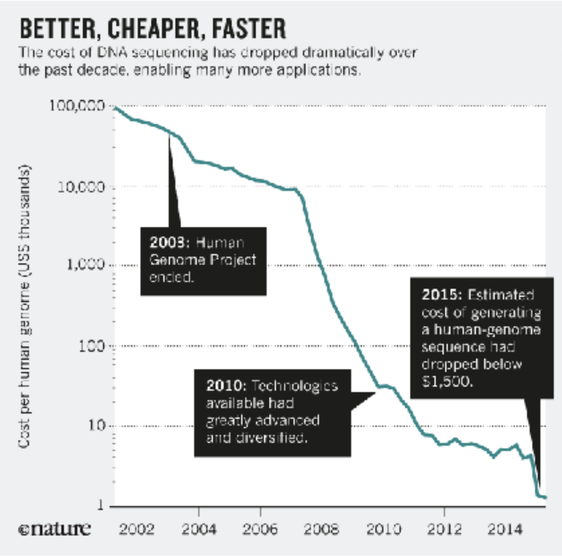
arrayMap

progenetix

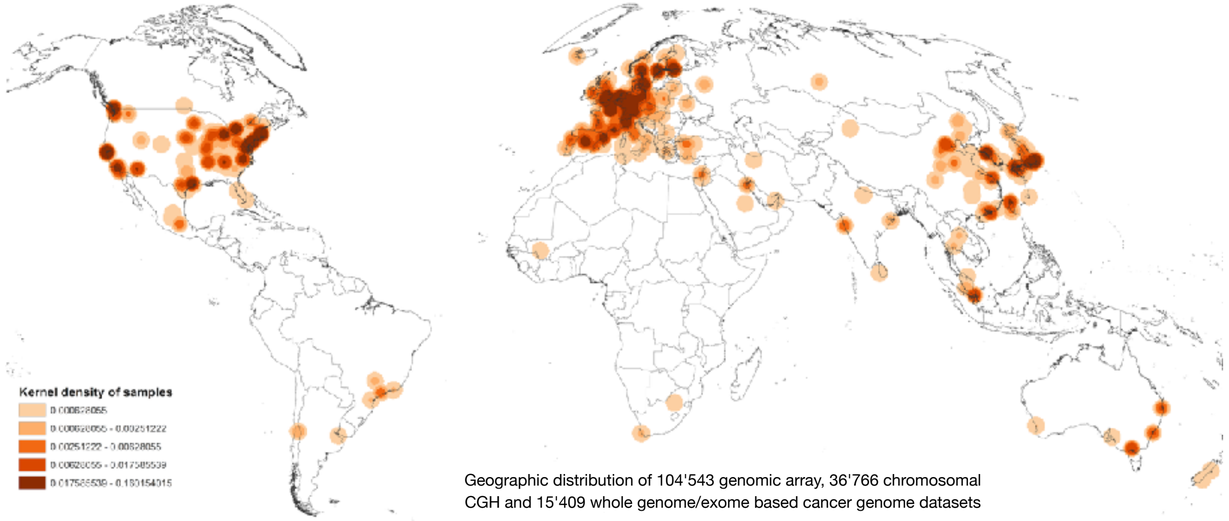
elixir

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)

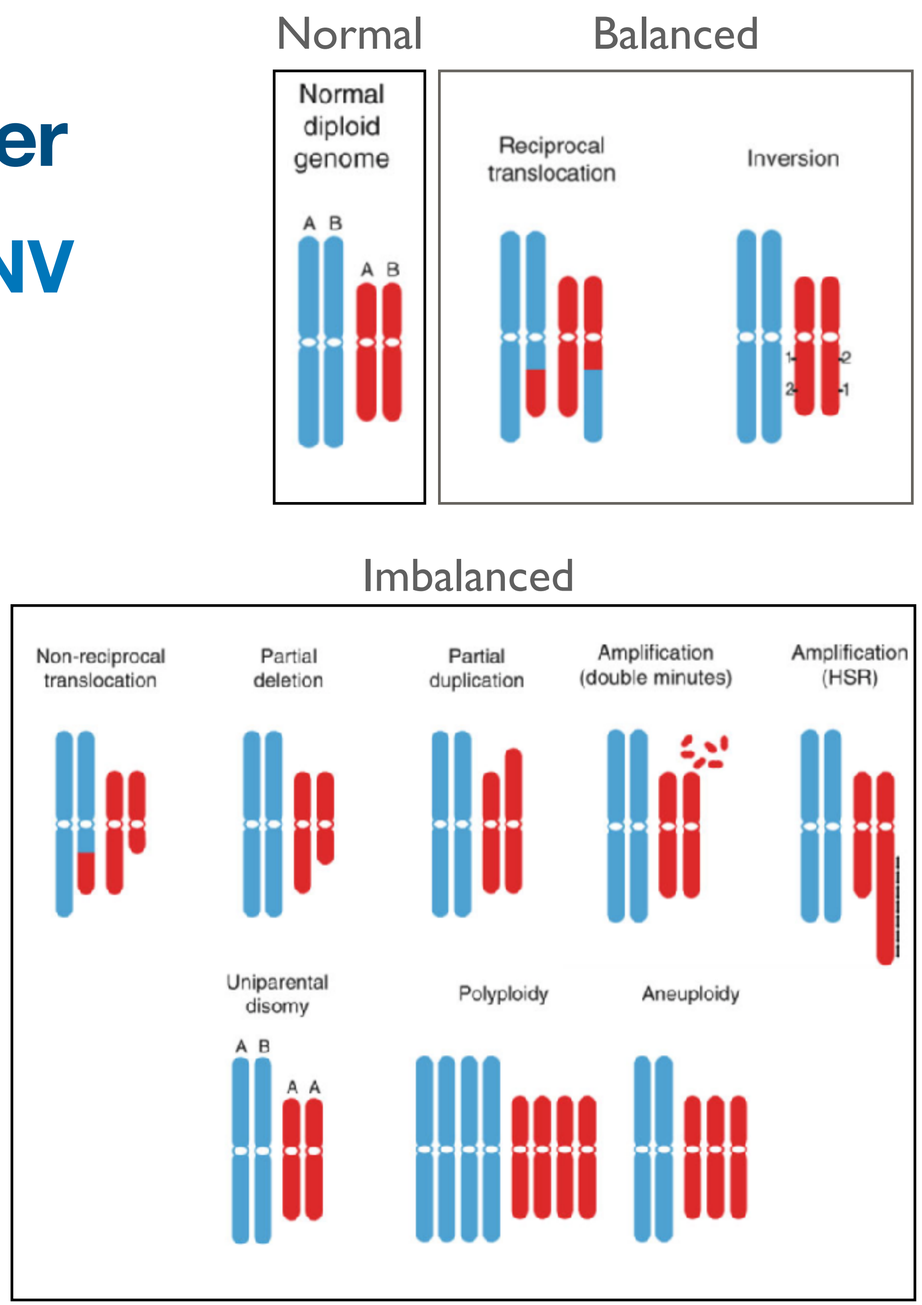


Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

Types of genomic alterations in Cancer

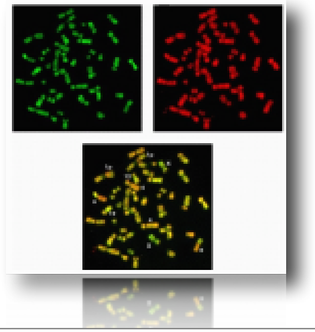
Imbalanced Chromosomal Changes: CNV

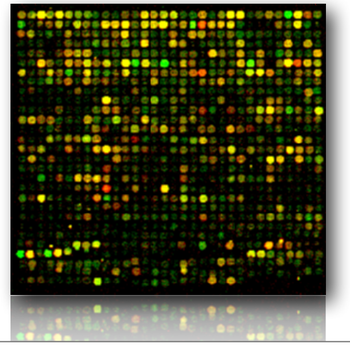
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- Structural chromosomal Aberrations
  - ➡ **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)

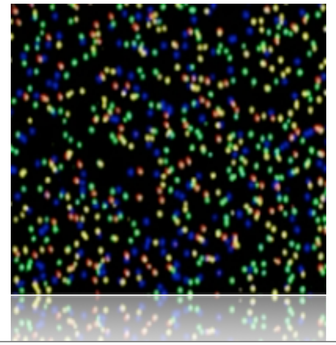


Grade et al., 2015 Recent Results Cancer Res

WHOLE GENOME SCREENING IN CANCER





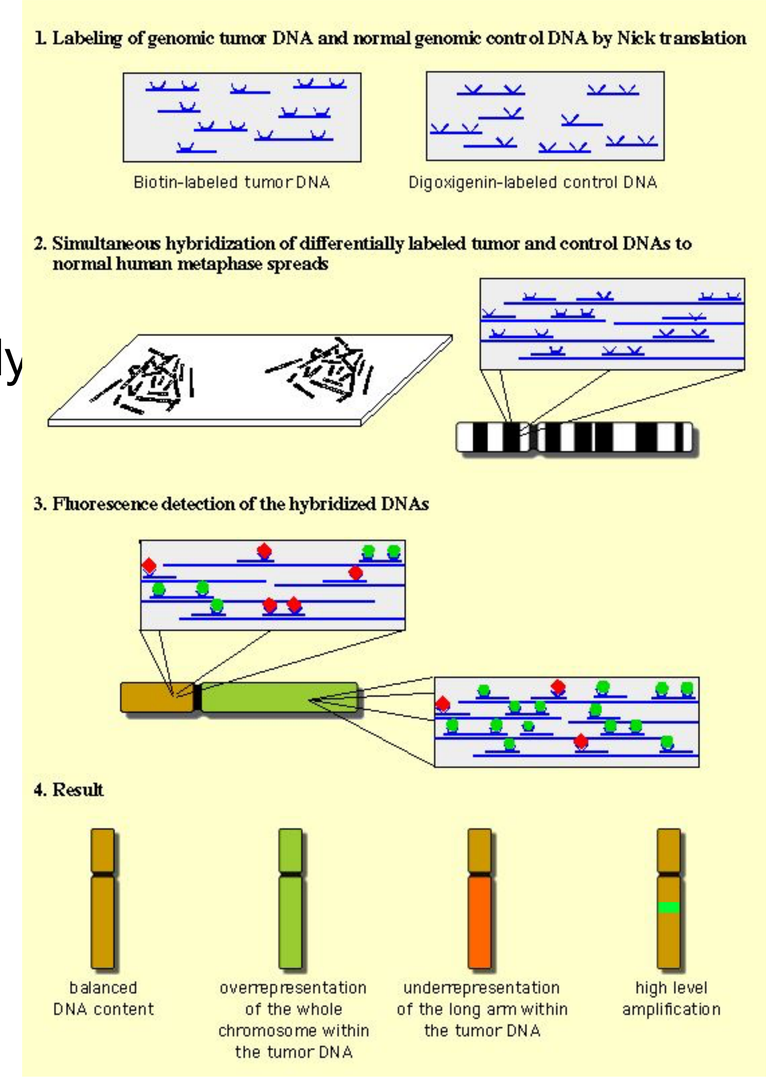


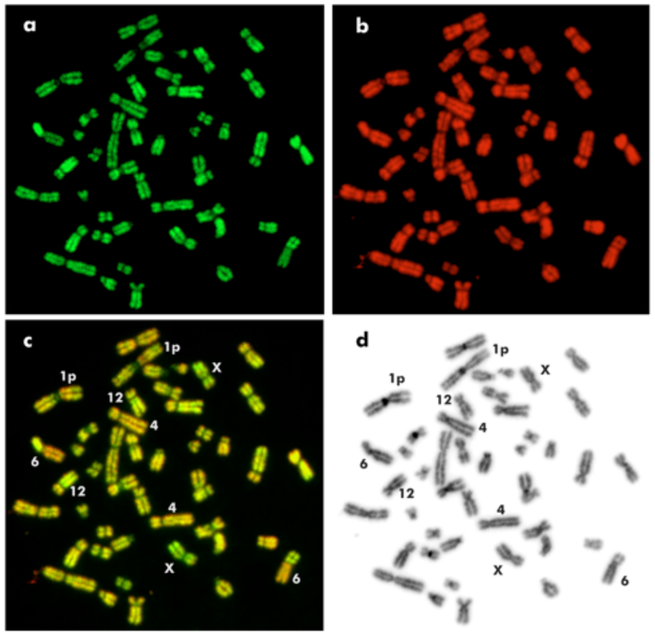
	Chromosomal CGH	Array CGH	"NGS" genome sequencing (WES, WGS)
1st application report	**1992**	**1997**	**2010**
source	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)
main source problems	mixed/degraded source tissue	mixed/degraded source tissue	mixed/degraded source tissue
resolution	chromosomal bands = few megabases	mostly in the 100kb range, but tiling possible	single bases
target identification	surrogate (position)	"semidirect" (segmentation spanning probes)	direct quantitative and qualitative
available data	>24,000 cases (57%) through **Progenetix**	raw data repositories (GEO, EMBL, SMD), **Progenetix**	Limited for raw data (BAMs ...); variant call data in dbgap, clinvar; selected studies with called CNV segments
predominant data format	ISCN = static	raw => depends on bioinformatics	mostly annotated variant calls or SNVs

Chromosomal Comparative Genomic Hybridization (CGH)

Molecular-Cytogenetic in situ hybridization

- Identify regional genomic copy number variations (CNV/CNA)
- **In situ hybridization** of **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 genes through cytogenetic bands (**megabase resolution**)





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

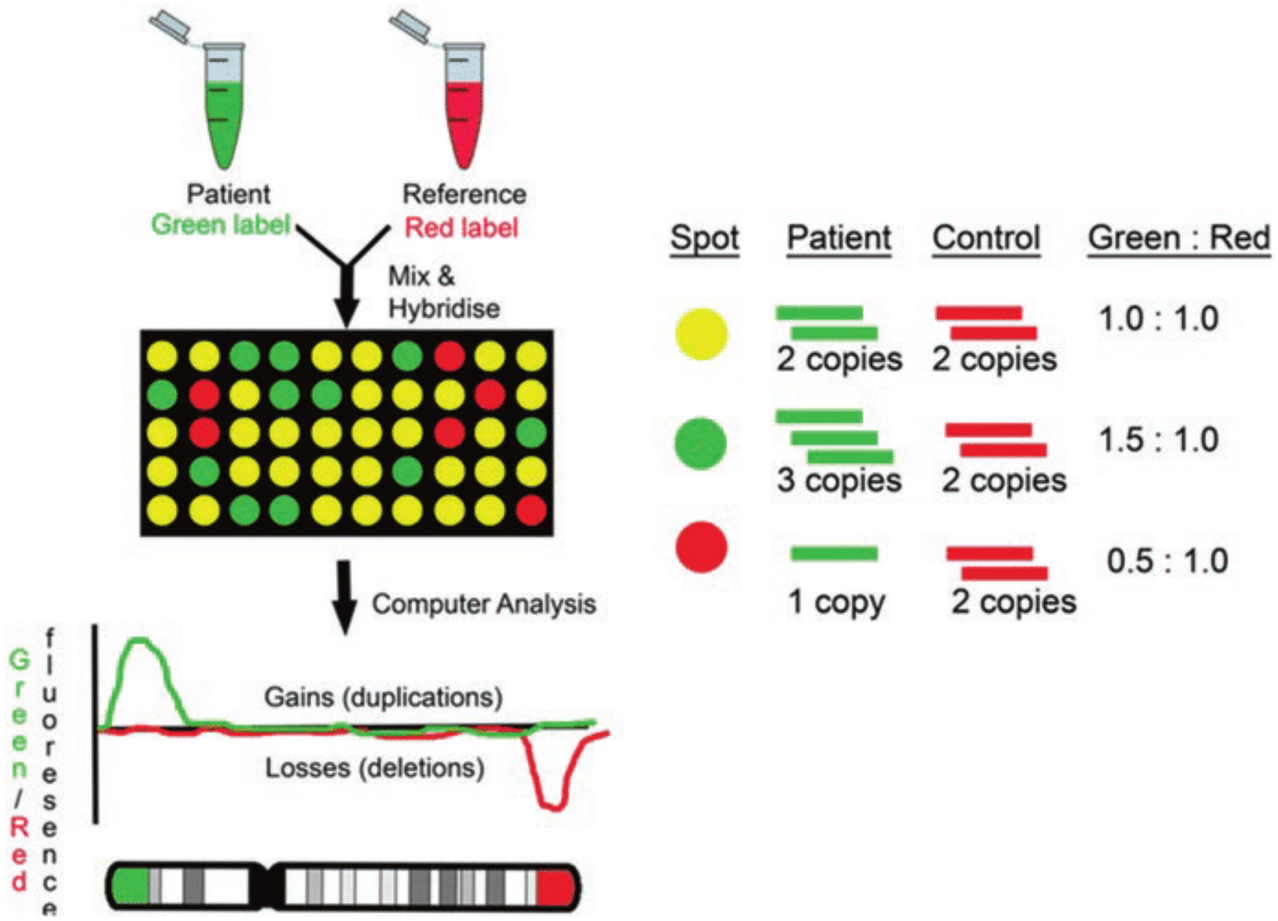
+6p, -6q...

- 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.

Array CGH

Fluorescent microarray with DNA probes

- Quantify ratio of probed DNA between patient and control samples
- Resolution ranges from **1 - 300 kb** on average depending on the platform
- Array probe design
  - cover clinically relevant locations
  - avoid repetitive sequences
  - distribute over whole range of genome

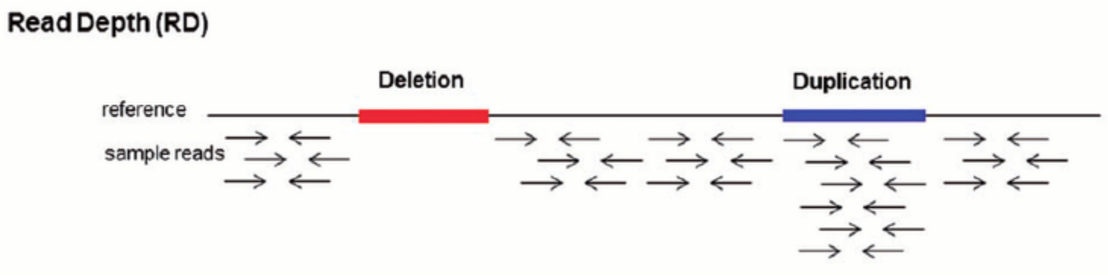


Techniques of Chromosomal Studies 2017 10.1007/978-81-322-3673-3_14

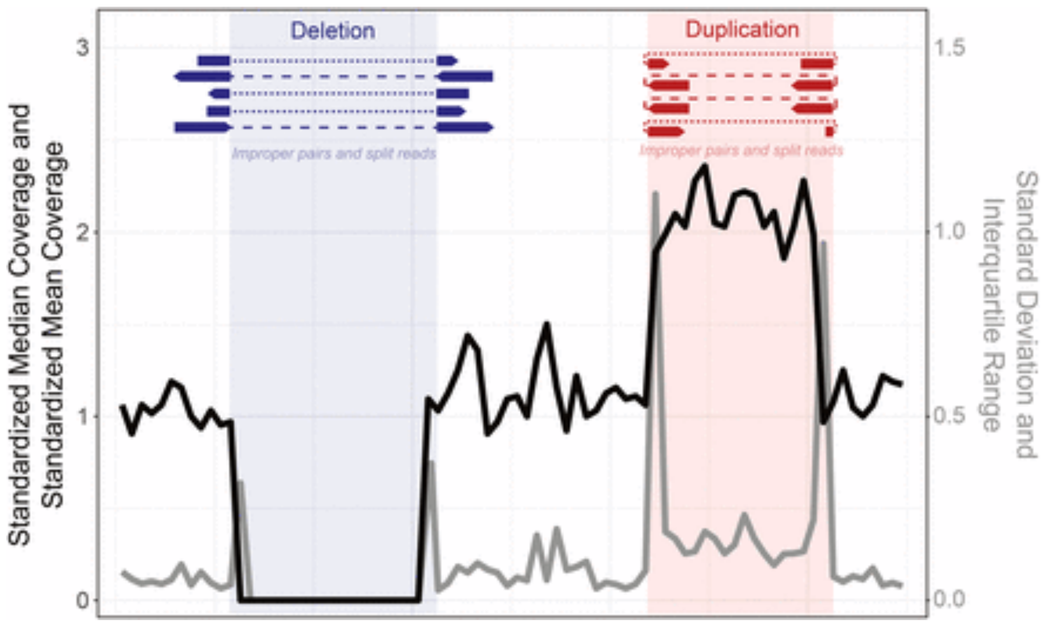
NGS-based method

WES, WGS

- Developed in 2010s
- Single base level detection
- Use read depth to quantify copy number change
- Possible to detect breakpoints
- Not directly standardized comparison, requires normalization

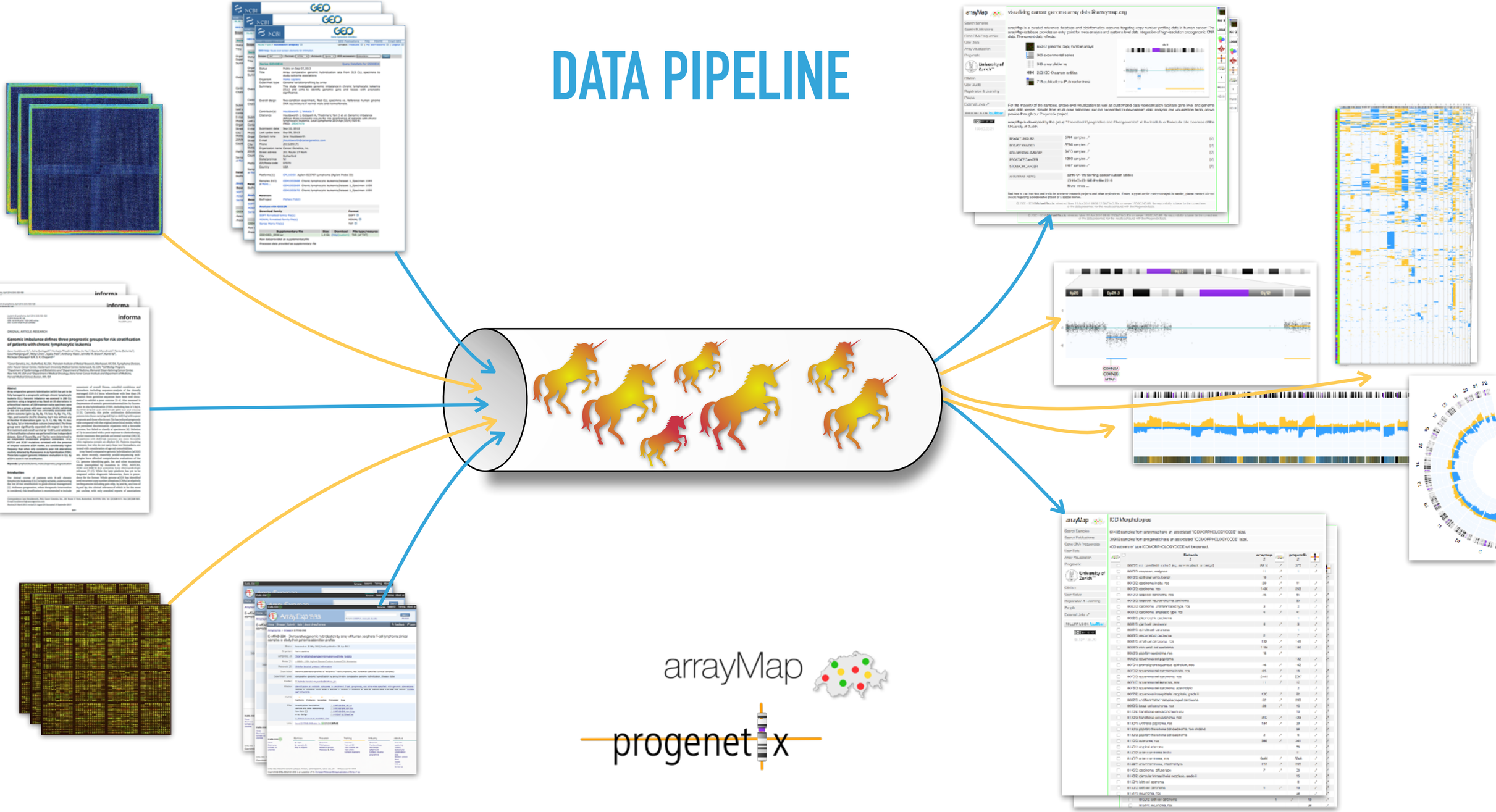


Sirbu et al., Apple Opt. 2016 10.1364/AO.55.006083

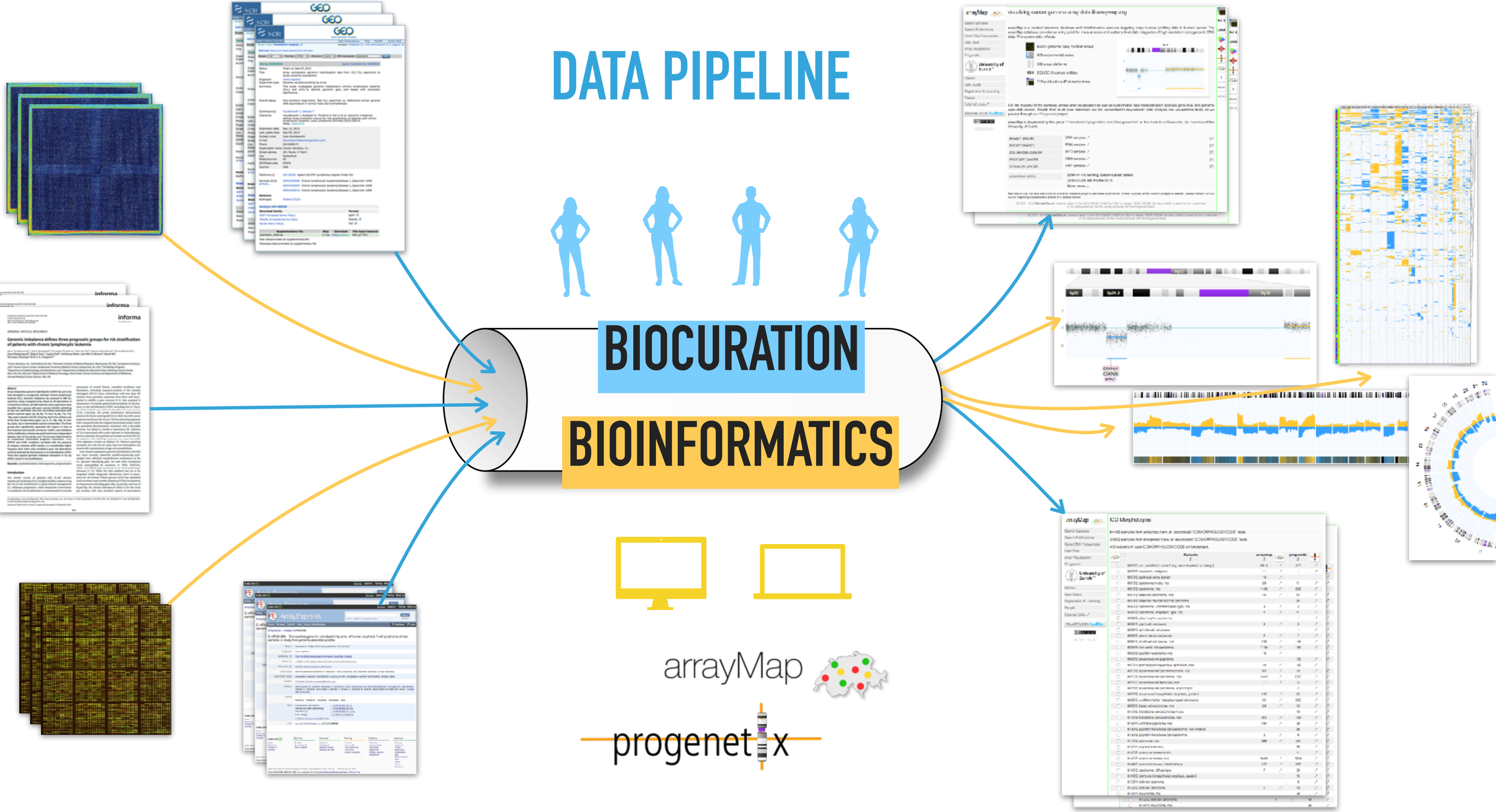


Hill and Uncles, 2019 G3 Genes|Genomes|Genetics 10.1534/g3.119.400596

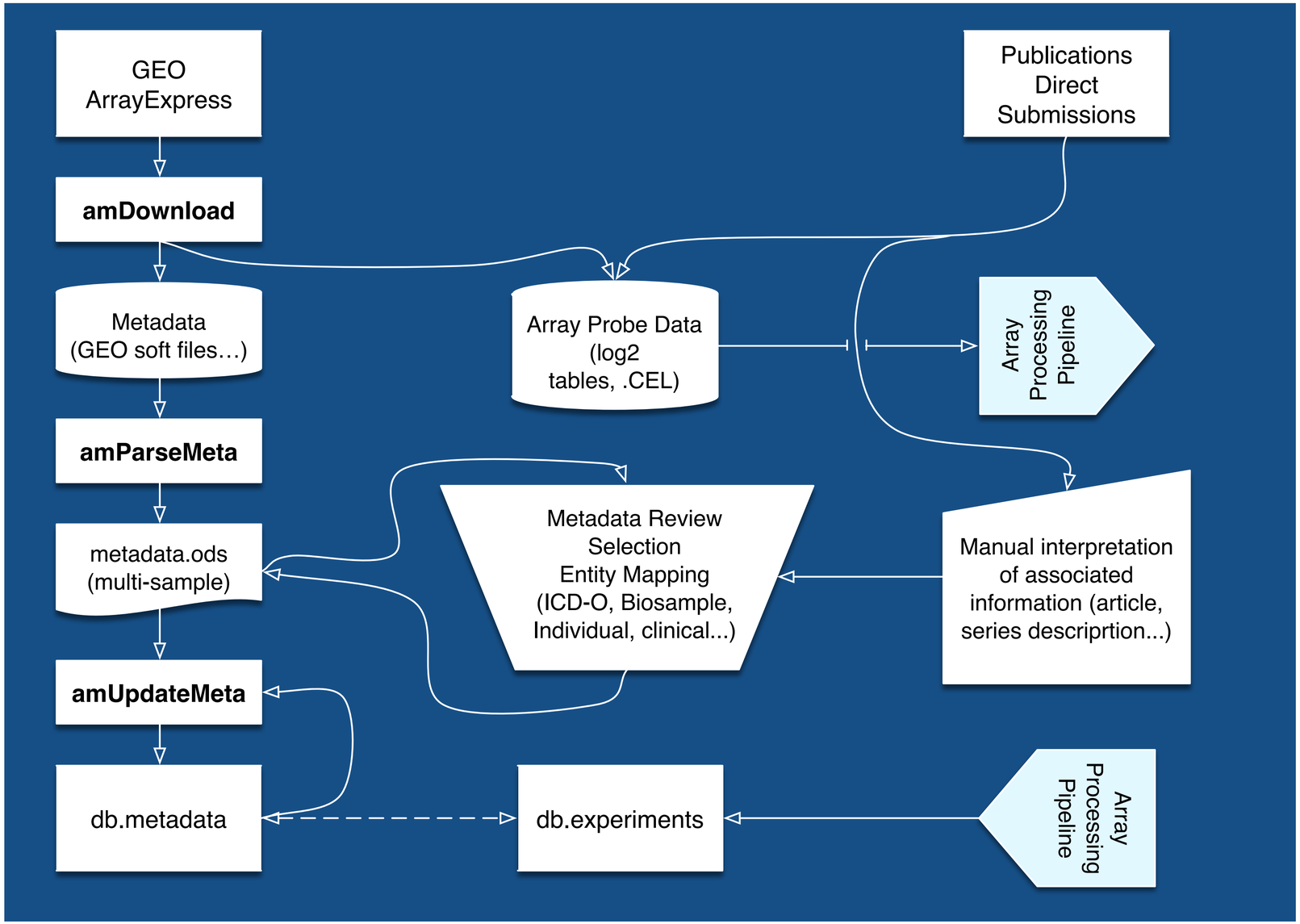
DATA PIPELINE



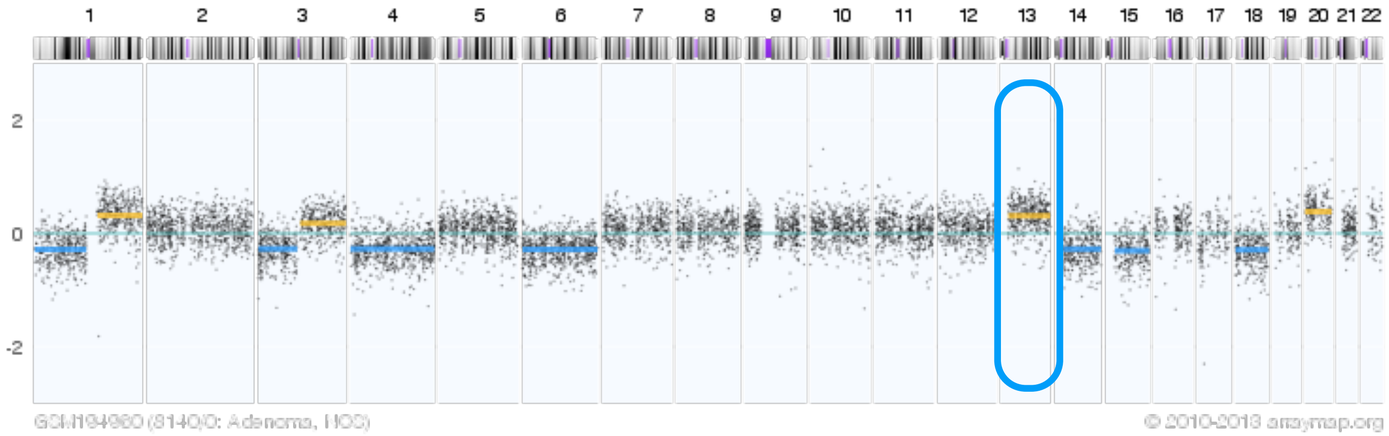
DATA PIPELINE



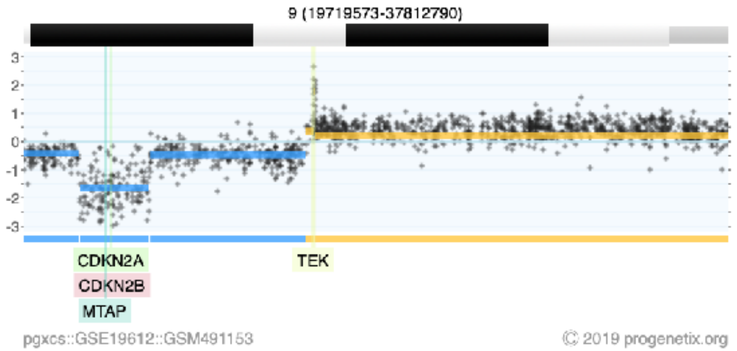
Bioinformatics & Data Curation - arrayMap data “Pipeline”



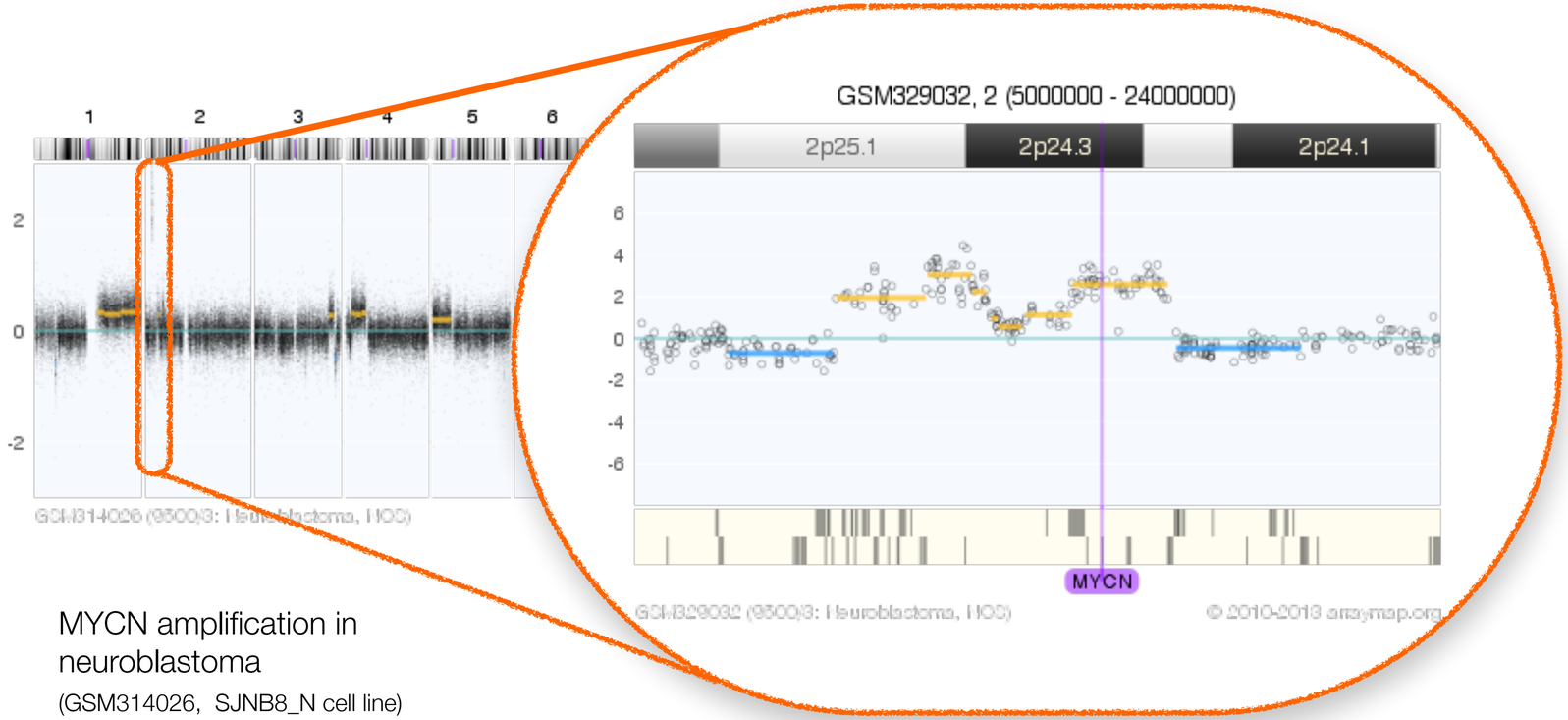
Somatic Copy Number Variations



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

What is Metadata?

- Summarize the data in a **structured, machine-readable** way.
- Describe the data using unique **identifiers**, and **controlled vocabularies**.
- **Searchable** in files, ontologies, websites and in registries.
- Essential to **Findable, Accessible, Interoperable and Reusable** (FAIR) bioinformatics.

Adapted from SIB course Metadata in Bioinformatics

Progenetix Metadata Scopes

Biomedical and procedural

- 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 = 8194
!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 = 8194
!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
701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard Affymetrix protocol from 2
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 ge
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

```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!

Source: https://xkcd.com/208/

```yaml
extraction_scopes:
  description: >-
    Detection and processing of clinical scopes goes through several stages:
    1. line cleanup - so far run for the input before processing the individual
    scopes
    2. line match using sme general pattern expected in all lines containing
    data for the current scope (`filter` pattern)
    3. finding and extracting the relevant data by looping over a list of
    specific patterns with memorized matches (`find`)
    4. post-processing using empirical cleanp replacements (`cleanup`)
    5. checking the correct structure (`final_check` - a global pattern can be
    used if other post-processing is performed)


  survival_status:
    filter: '(?i).*?(?:(?:dea(?:d|th))|alive|surviv|outcome|status)'
    preclean:
      - m: '(?i)days to death or last seen alive[^\w]+?\d+?(?:[^\w\.]|$)'
        s: ''
      - m: '[^\w]+?NA(?:[^\w\.]|$)'
        s: ' '
      - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?ED'
        s: 'survival: dead'
      - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?NA'
        s: ''
      - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?CR'
        s: 'survival: alive'
      - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?RD'
        s: '' # alive but not responding to therapy so removed?
      - m: 'Event Free Survival[^\w]+?no event'
        s: 'recurrence: no'
      - m: 'Event Free Survival.event'
        s: 'recurrence: yes'
      - m: 'Outcome[^\w]+?no event'
        s: 'survival: alive'
      - m: 'Outcome[^\w]+?event'
        s: 'survival: dead'
      - m: 'survival status[^\w]+?0'
        s: 'survival: dead'
      - m: 'survival status[^\w]+?1'
        s: 'survival: alive'
      - m: 'overall[^\w]+?survival[^\w]+?days[^\w]+?NA'
        s: ''
      - m: 'survival(?: time|from diagnosis)?[^\w]+?(days|months|years?)[^\w]+?(\d\d?\d?\d?\.?\d?\d?)'
        s: 'survival: \2\1'
```

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

- Historically classified using the 2 arms of the ICD-O system
  - morphology ~ histology
  - topography ~ organ/tissue
- mappings between ICD-O code pairs and the NCIt "neoplasm" part of the NCI meta-thesaurus empower **hierarchical** data structures for search and analysis

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 ...)

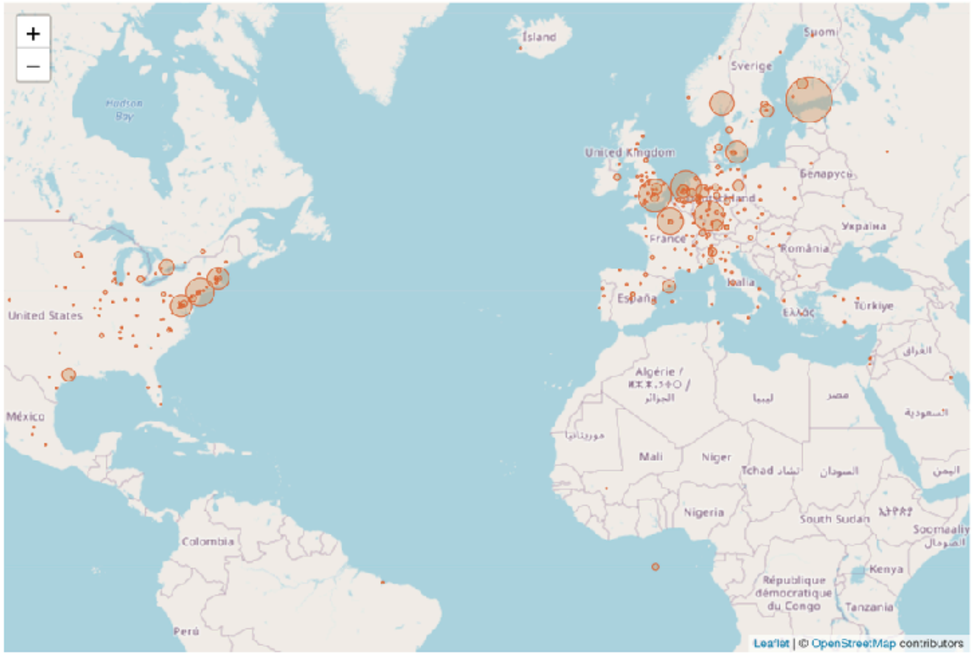

```
"data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
},

"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!

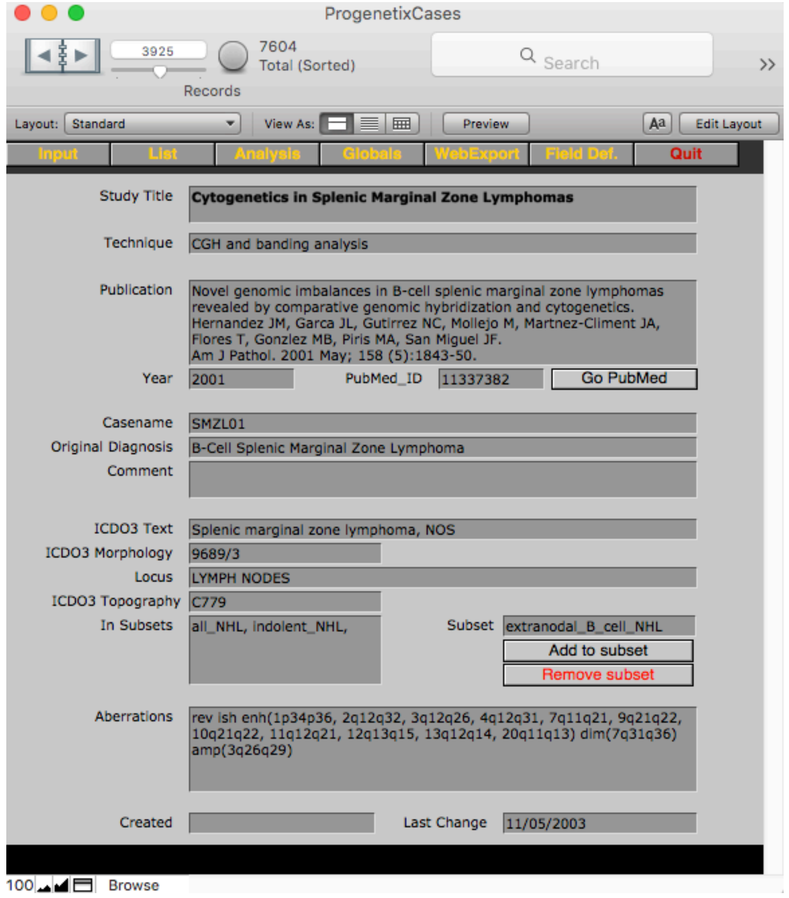


Geographic distribution (by corresponding author) of the **118554** genomic array, **36766** chromosomal CGH and **42105** whole genome/exome based cancer genome datasets from the **3306** listed publications. Area sizes correspond to the sample numbers reported from a given location.

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

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

2024

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

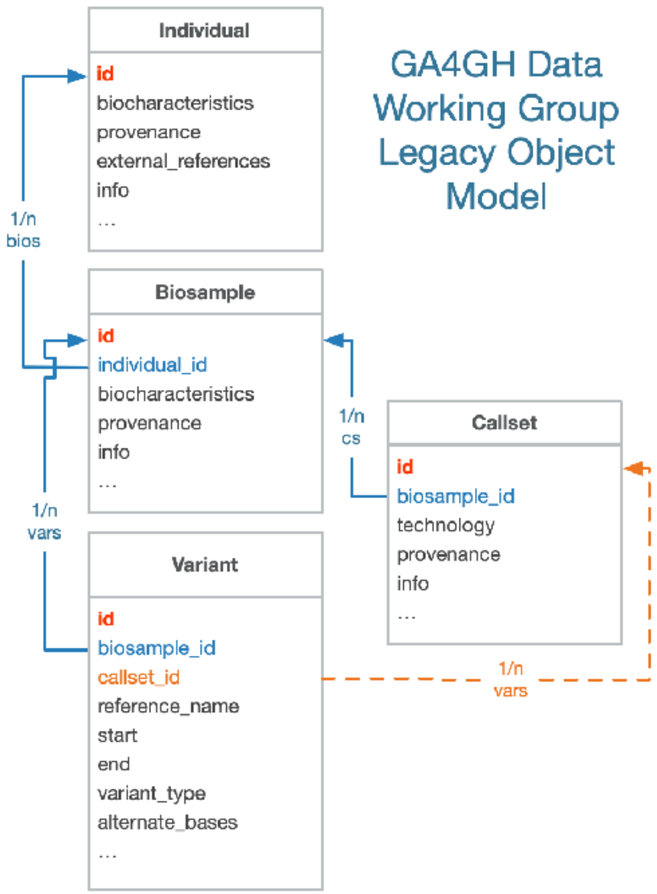

```
{
 "id" : "pgxind-kftx394x",
 "biocharacteristics" : [
  {
   "description" : "female",
   "type" : {
    "id" : "PATO:0020002",
    "label" : "female genotypic sex"
   }
  },
  {
   "description" : null,
   "type" : {
    "id" : "NCBITaxon:9606",
    "label" : "Homo sapiens"
   }
  }
 ],
 "data_use_conditions" : {
  "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



- 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**

2024

- 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"
 }
}
```

Progenetix in 2024

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiles
- focus on copy number variations (CNV)
- >116'000 cancer CNV profiles, from >800 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCIt, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

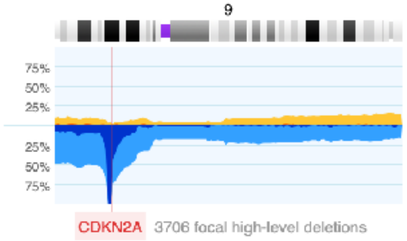
progenetix

Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* of currently **136468** samples from **834** different cancer types (NCIt neoplasm classification)

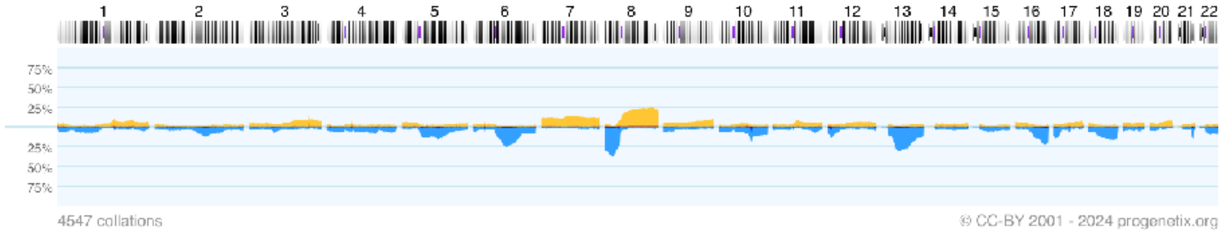
Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [Search Page] provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.



Cancer CNV Profiles

Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the respective Cancer Types pages with visualization and sample retrieval options. Below is a typical example of the aggregated CNV data in 4547 samples in Malignant Male Reproductive System Neoplasm with the frequency of regional copy number gains (**high level**) and losses (**high level**) displayed for the 22 autosomes.



Download SVG | Go to NCIT:C8561 | Download CNV Frequencies

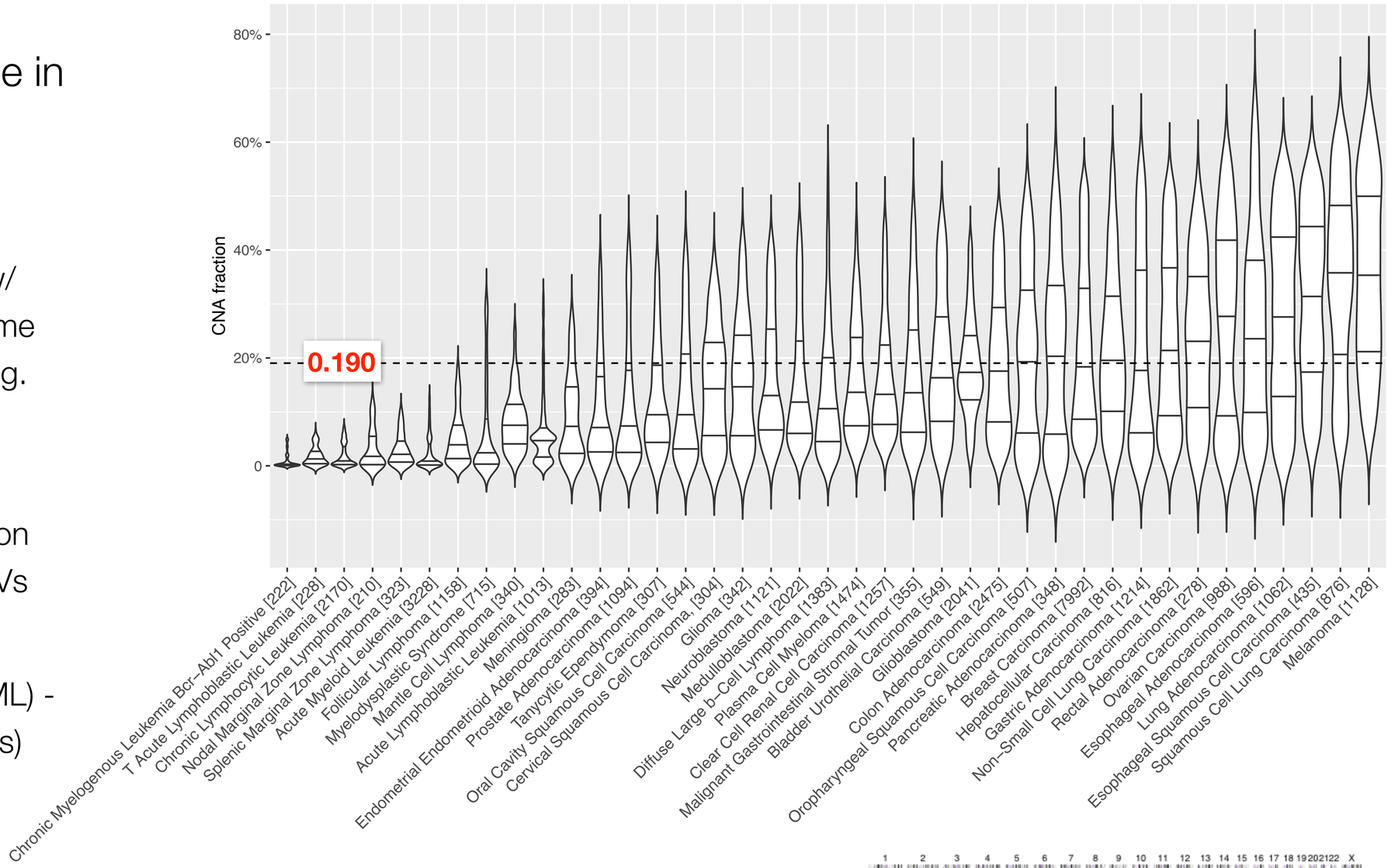
Cancer Genomics Publications

Through the [Publications] page Progenetix provides annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

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)

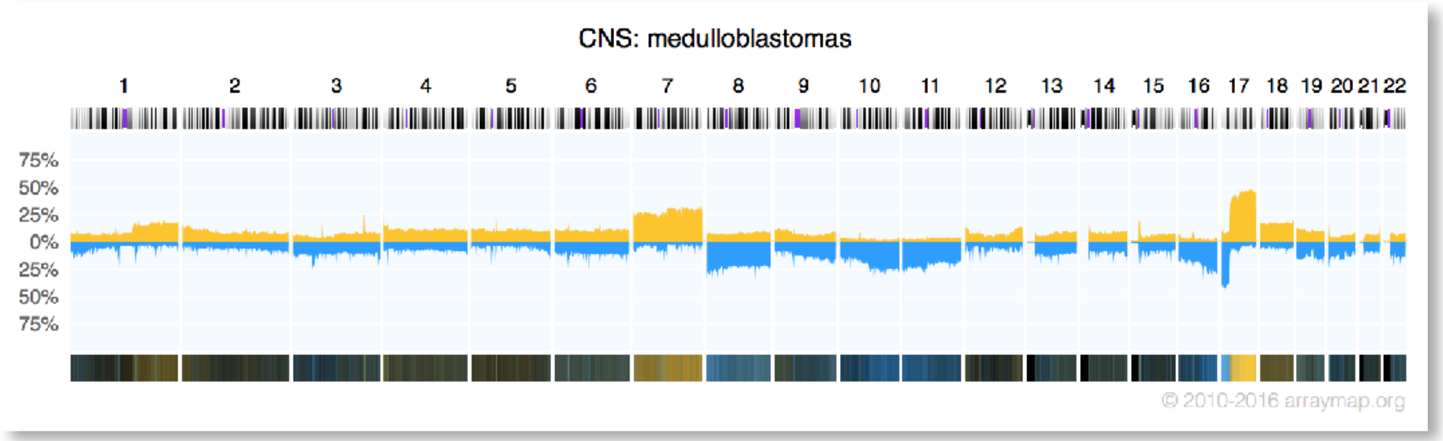


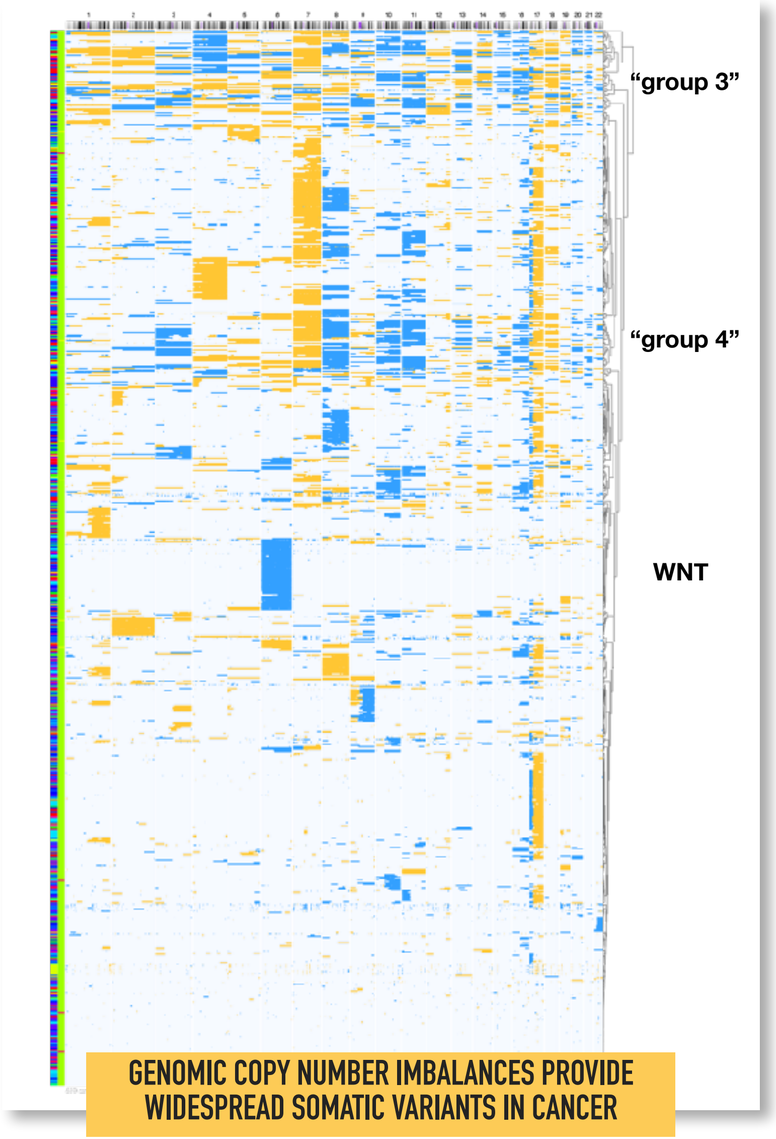
Lowest / Highest CNV fractions =>

Somatic CNVs In Cancer

Recurrent mutation patterns

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





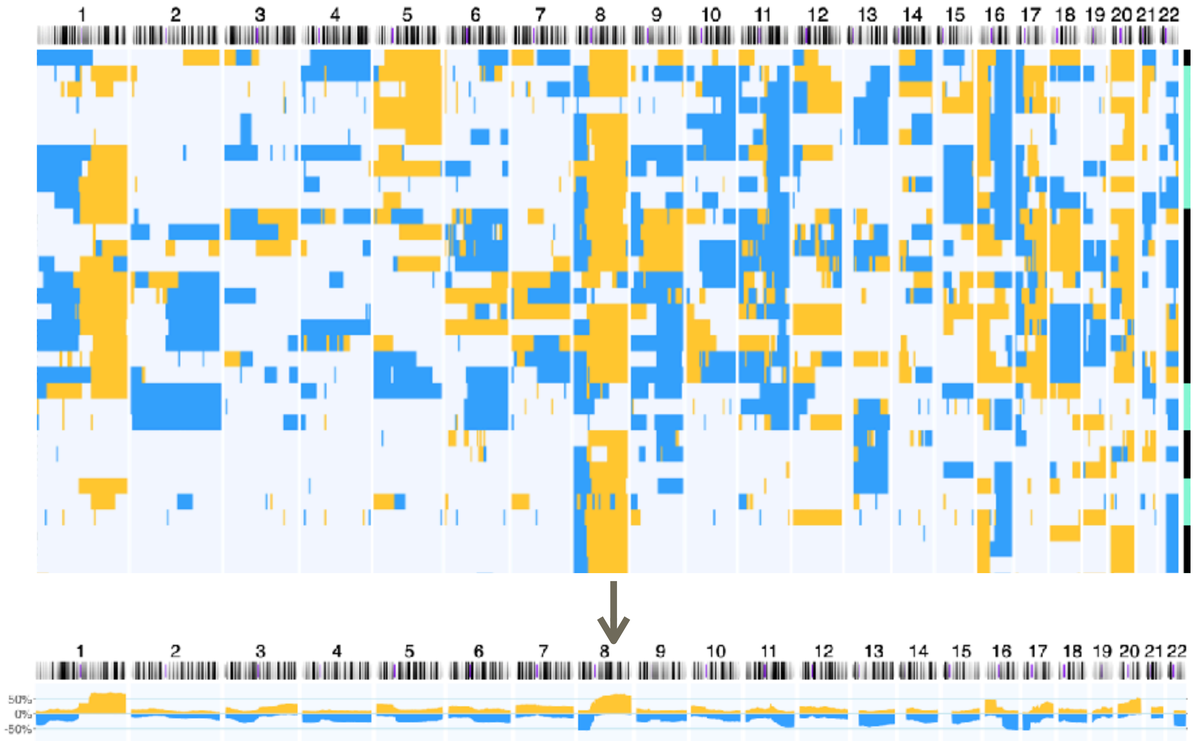
A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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

arrayMap

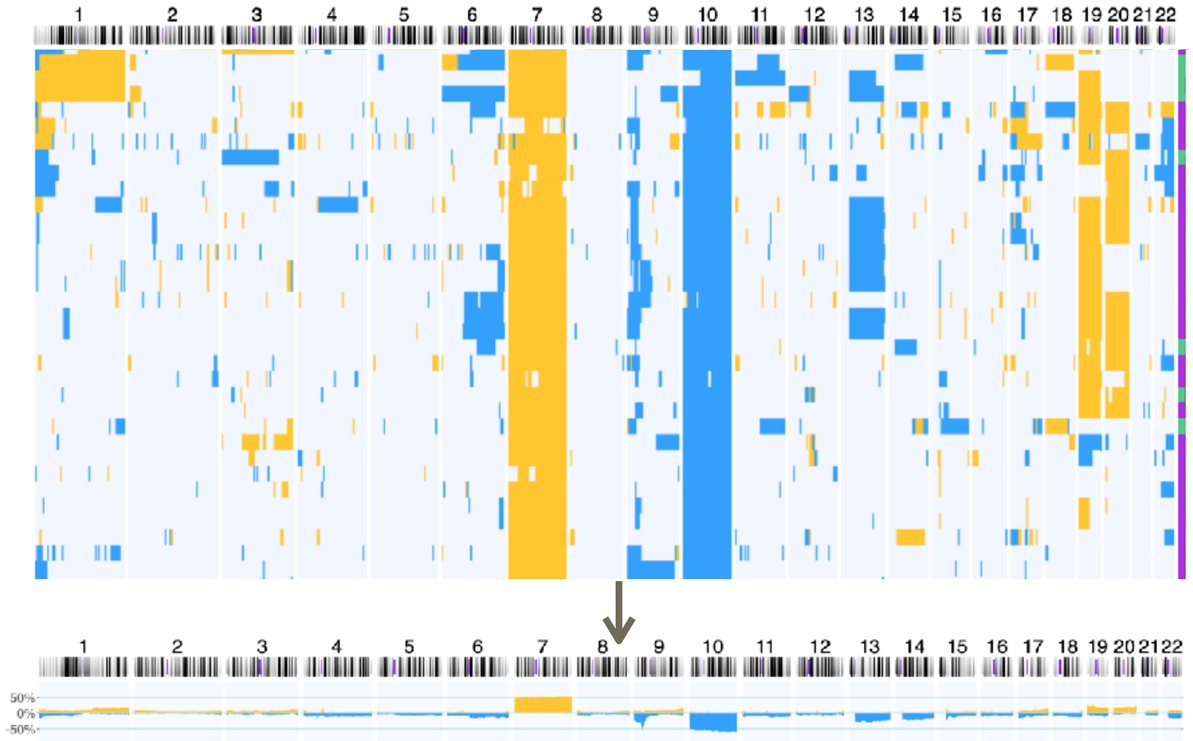
Drivers? Passengers? Markers?

Disentangling CNA Patterns

Ductal Breast Carcinoma



Glioblastoma



Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles

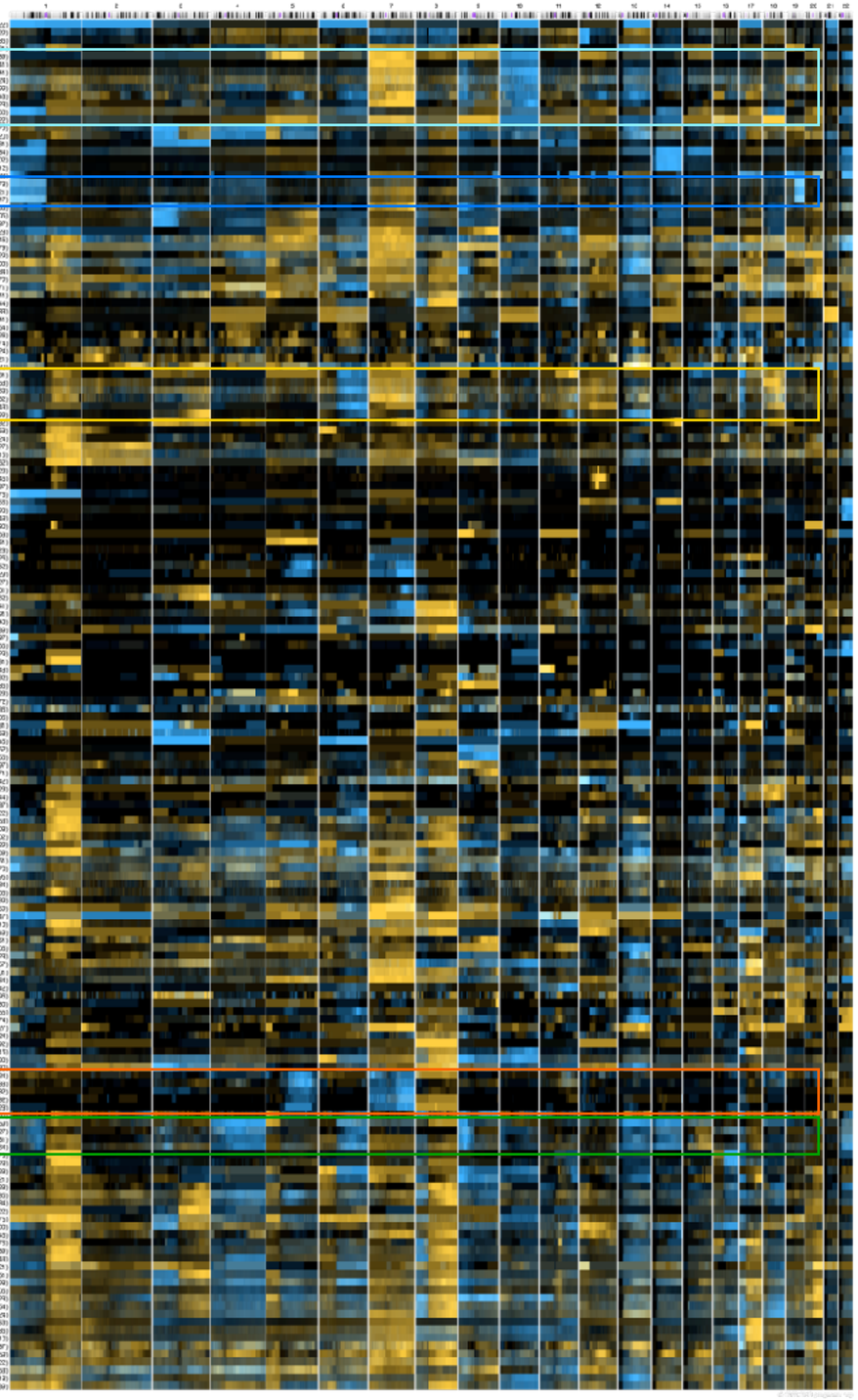
9390/1: choroid plexus papilloma, nos (39)
9442/3: gliosarcoma (41)
9440/3: glioblastoma, nos (1241)
9401/3: astrocytoma, anaplastic (124)
9380/3: glioma, nos (99)
9702/3: malignant lymphoma, t-cell nos (48)
9381/3: gliomatosis cerebri (23)
9530/3: meningioma, malignant (60)
9394/1: myxopapillary ependymoma (22)

9451/3: oligodendroglioma, anaplastic (78)
9382/3: oligoastrocytoma (121)
9450/3: oligodendroglioma, nos (147)

9698/3: follicular lymphoma, grade 3 (31)
9690/3: follicular lymphoma, nos (753)
9680/3: diffuse large b-cell lymphoma, nos (1263)
9591/3: malignant lymphoma, b-cell nos (62)
9590/3: malignant lymphoma, nos (43)
9673/3: mantle cell lymphoma (499)

9984/3: refractory anemia with excess blasts in transformation [raebt] (24)
9983/3: refractory anemia with excess blasts [raeb] (38)
9867/3: acute myelomonocytic leukemia [fab type m4] (32)
9920/3: therapy-related acute myeloid leukemia, nos (32)
9891/3: acute monoblastic leukemia [fab m5] (23)

9051/3: desmoplastic mesothelioma (59)
9053/3: mesothelioma, biphasic, malignant (27)
9050/3: mesothelioma, nos (81)
9052/3: epithelioid mesothelioma, malignant (64)



arrayMap

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fgene.2021.654887

Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes

*Bo Gao [1,2] and Michael Baudis [1,2]**

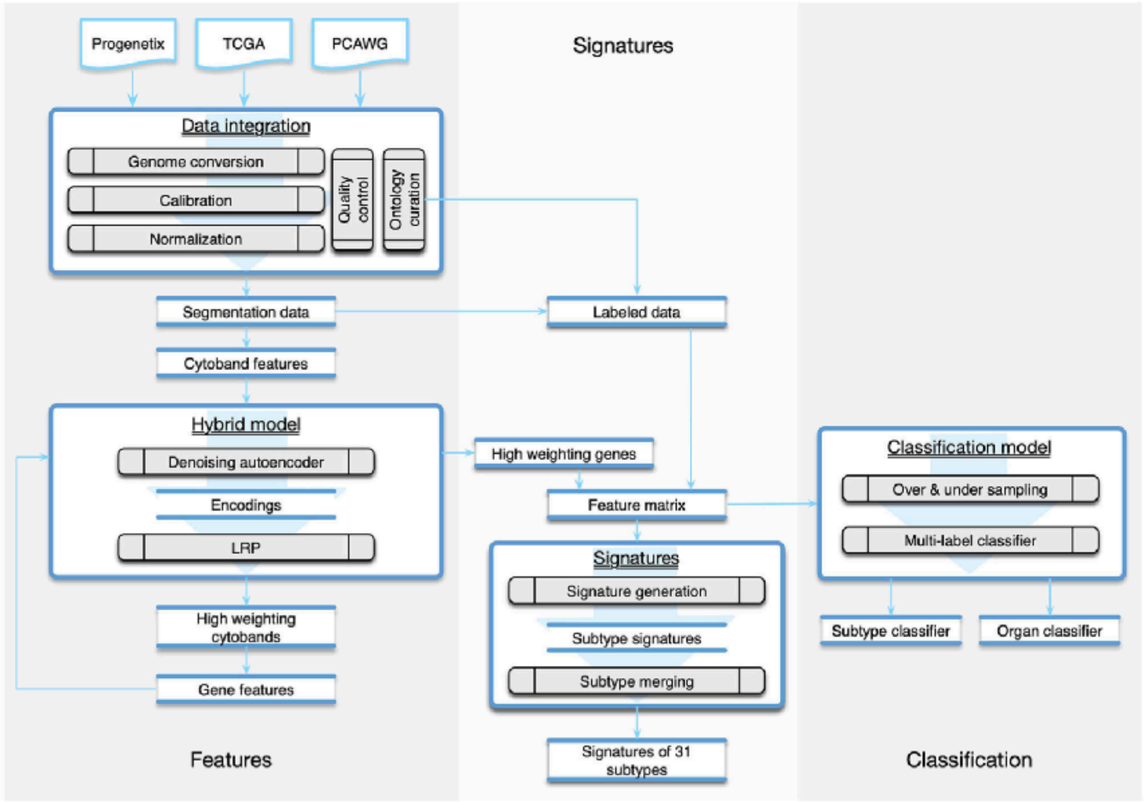


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine learning techniques to predict the organ and the subtype from a given copy number profile.

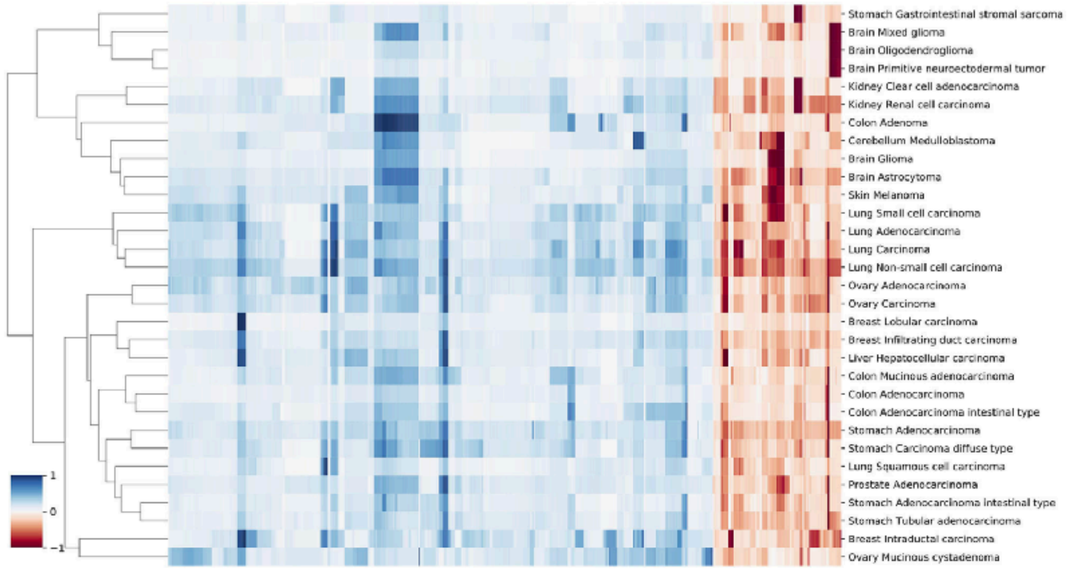


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

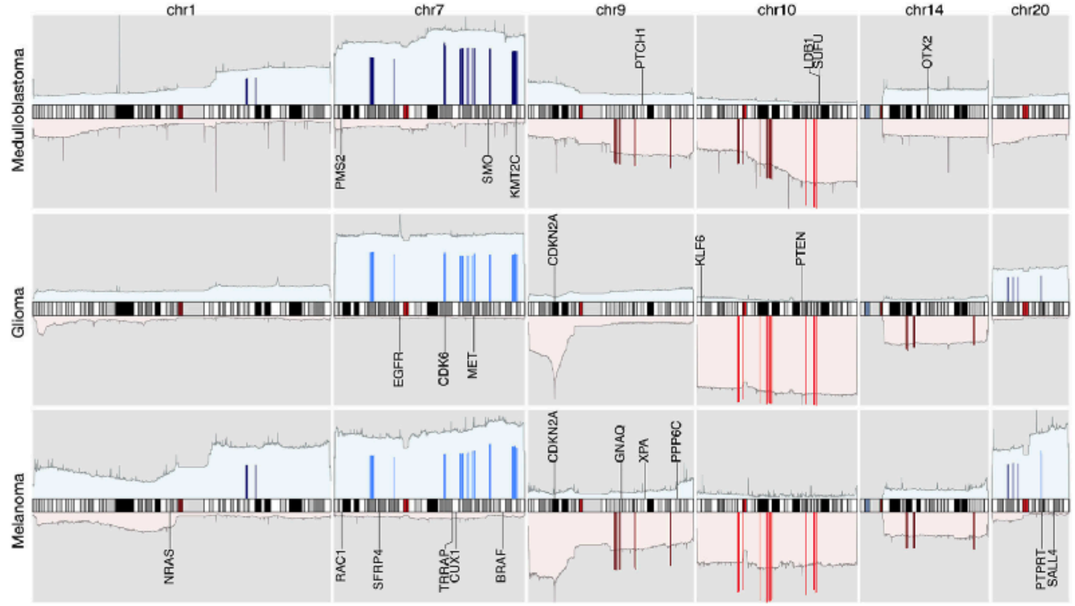


FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger signal intensity. The blue colors above the chromosome axis represent the average amplifications, and the red colors below the chromosome axis represent the average deletions. The amplitude of amplifications and deletions are normalized to [0, 1] separately. The adjacent known driver genes are also included for each tumor type.

Population stratification in cancer samples based on SNP array data

- Despite extensive somatic mutations of cancer profiling data, consistency between germline and cancer samples reached 97% and 92% for 5 and 26 populations
- Comparison of our benchmarked results with self-reported meta-data estimated a matching rate between 88 % to 92%.
- Ethnicity labels indicated in meta-data are vague compared to the standardized output from our tool

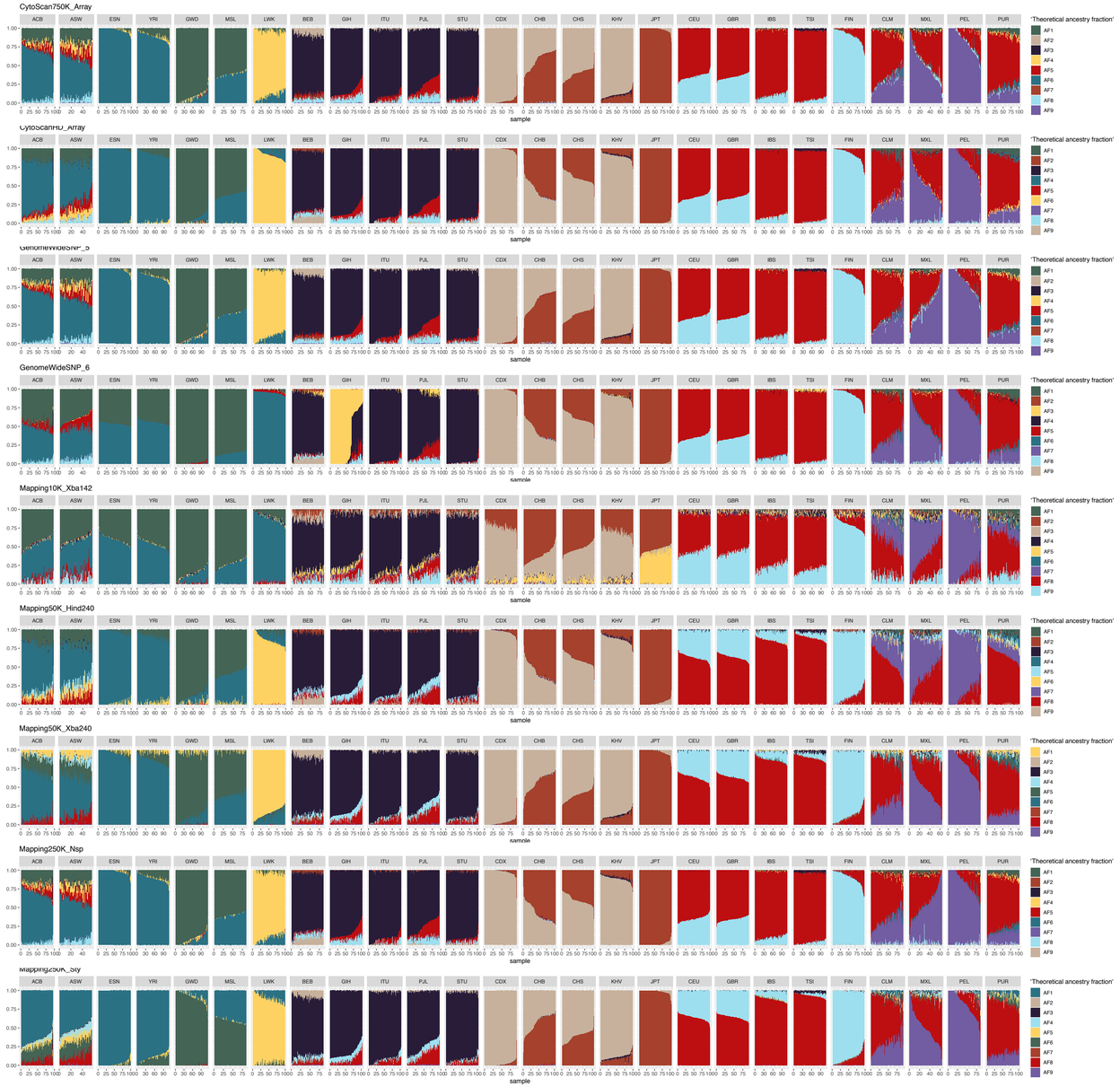


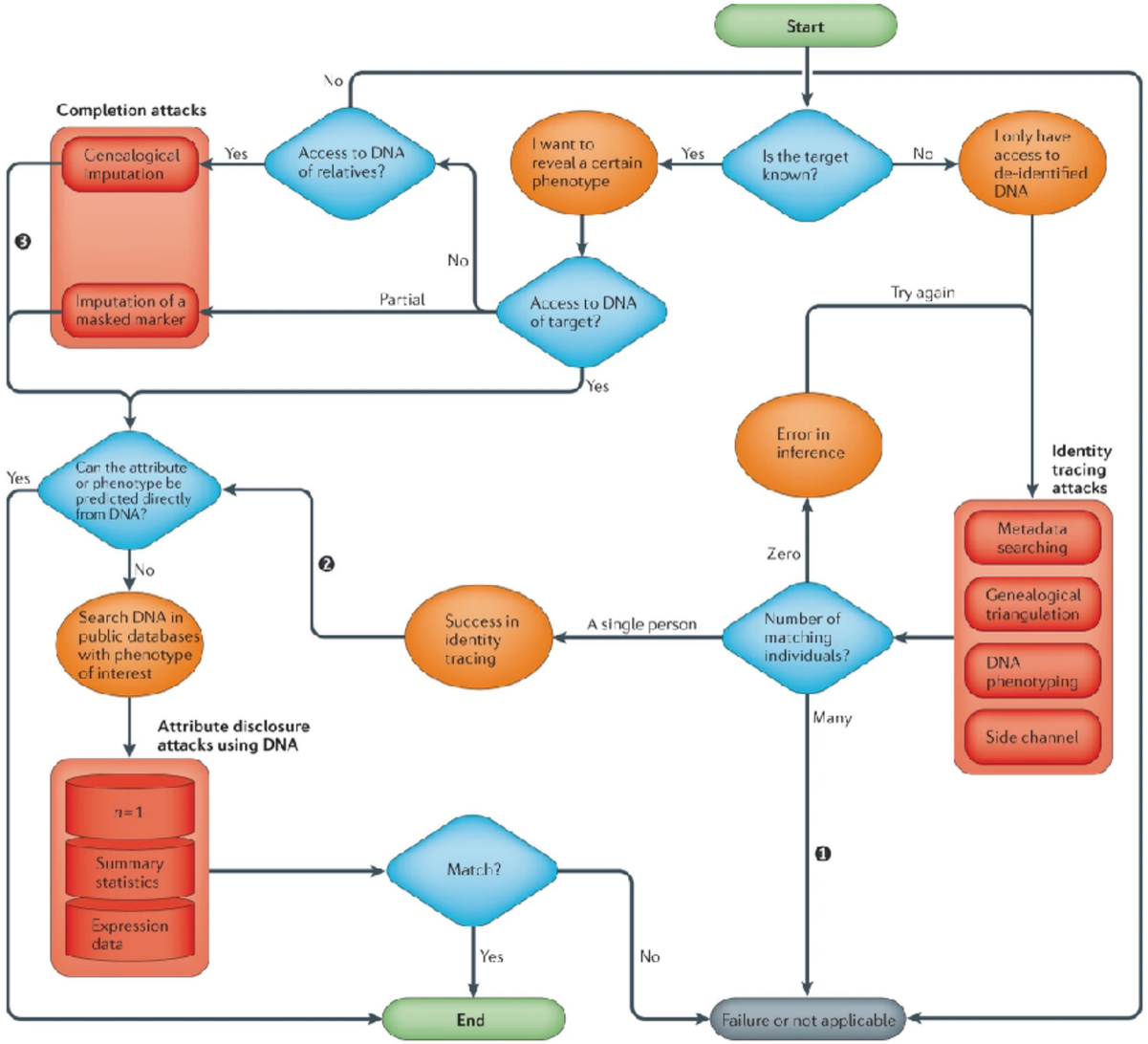
Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed neighboring to each other: AFR (1-7), SAS (8-12), EAS (13-17), EUR (18-22), AMR (23-26).

arrayMap

Qingyao Huang

How to share patient data safely?

- Risks
  - Long-range familiar search
    - “Golden State killer” - Cold cases in 1970s
    - DNA evidence led to capture in 2018
  - Membership inference attack
  - Reconstruction attack
- Privacy by design
  - Access control
  - Data aggregation
  - Data obfuscation



Wan et al. 2022 Sociotechnical safeguards for genomic data privacy. NRG

Erlich and Narayanan 2014. Routes for breaching and protecting genetic privacy. NRG

GENOMICS

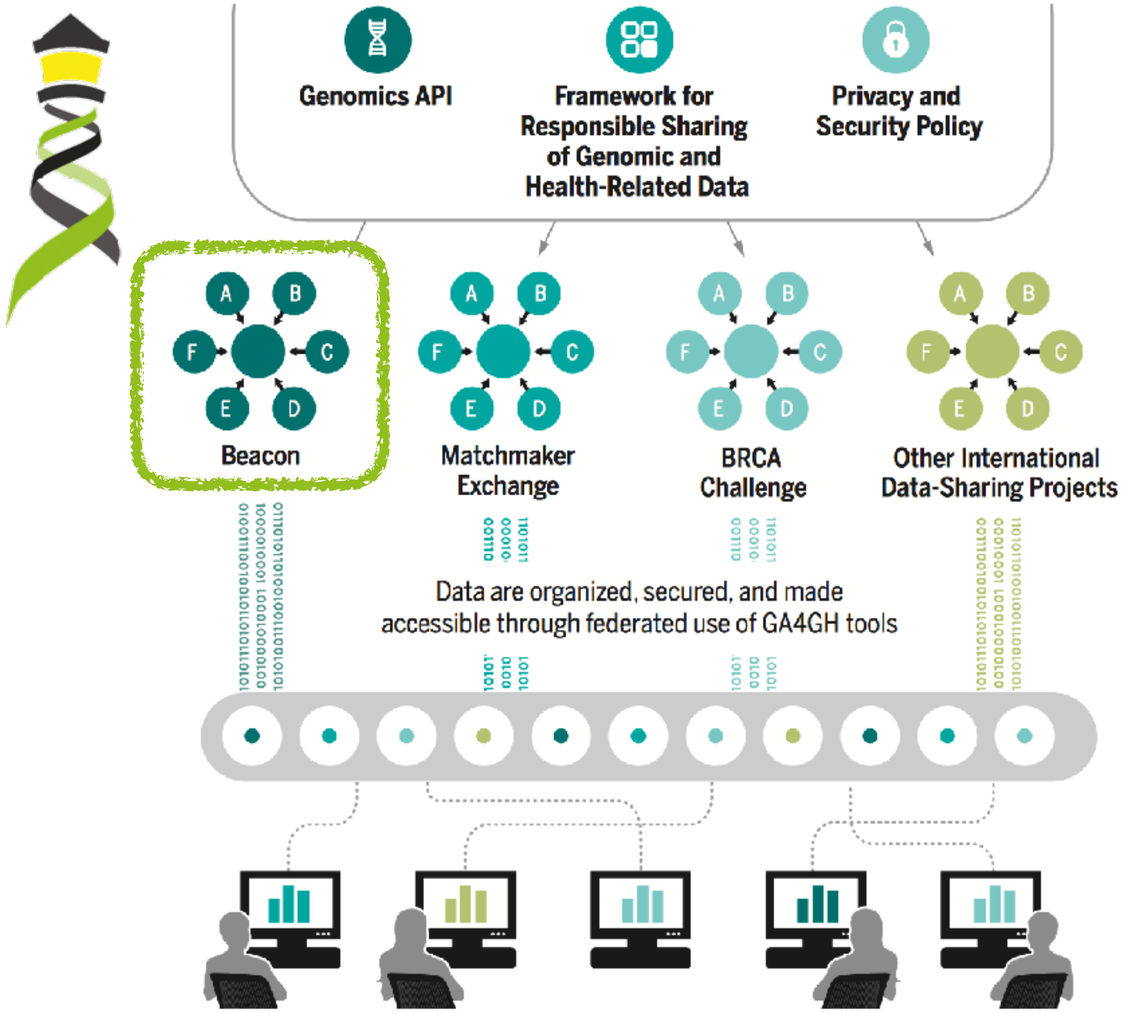
A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



DNASTACK

Global Alliance for Genomics & Health

Progenetix and GA4GH Beacon

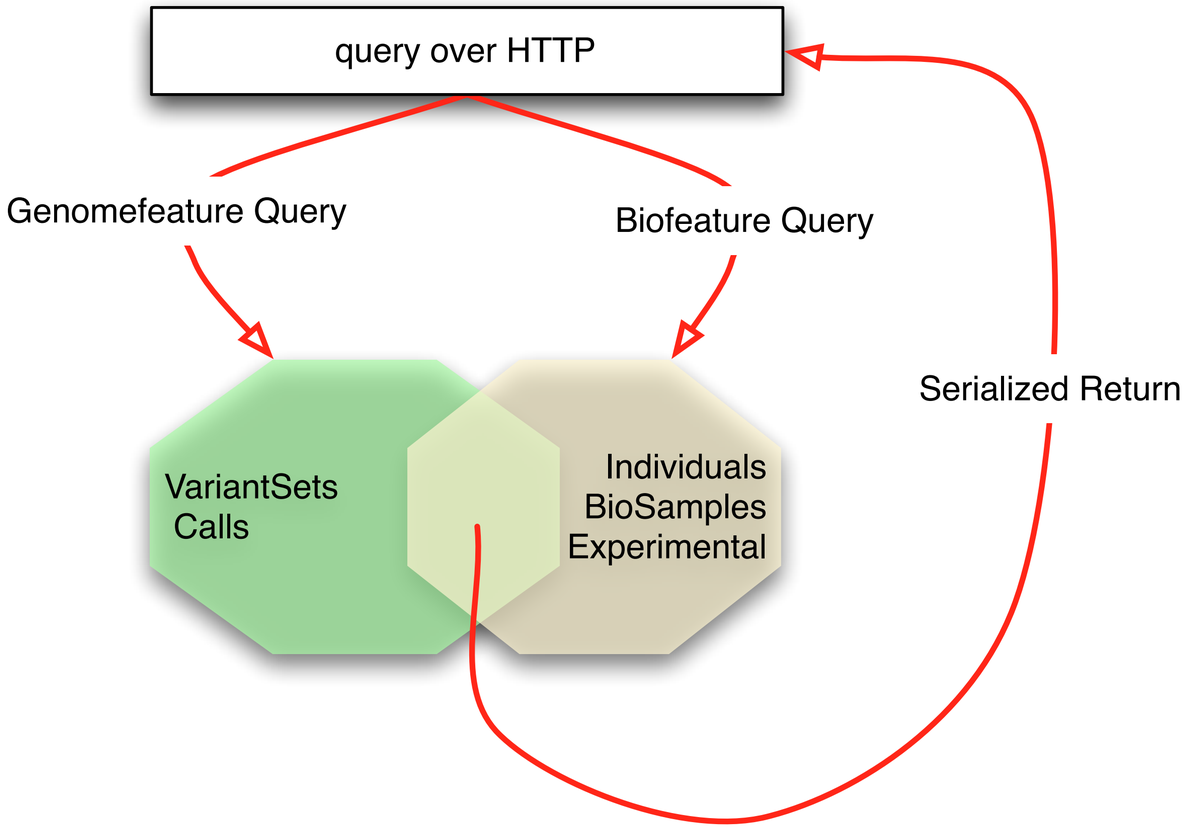
Implementation driven development of a GA4GH standard

progenetix

A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

YES | **NO** | **\0**

Minimal GA4GH query API structure



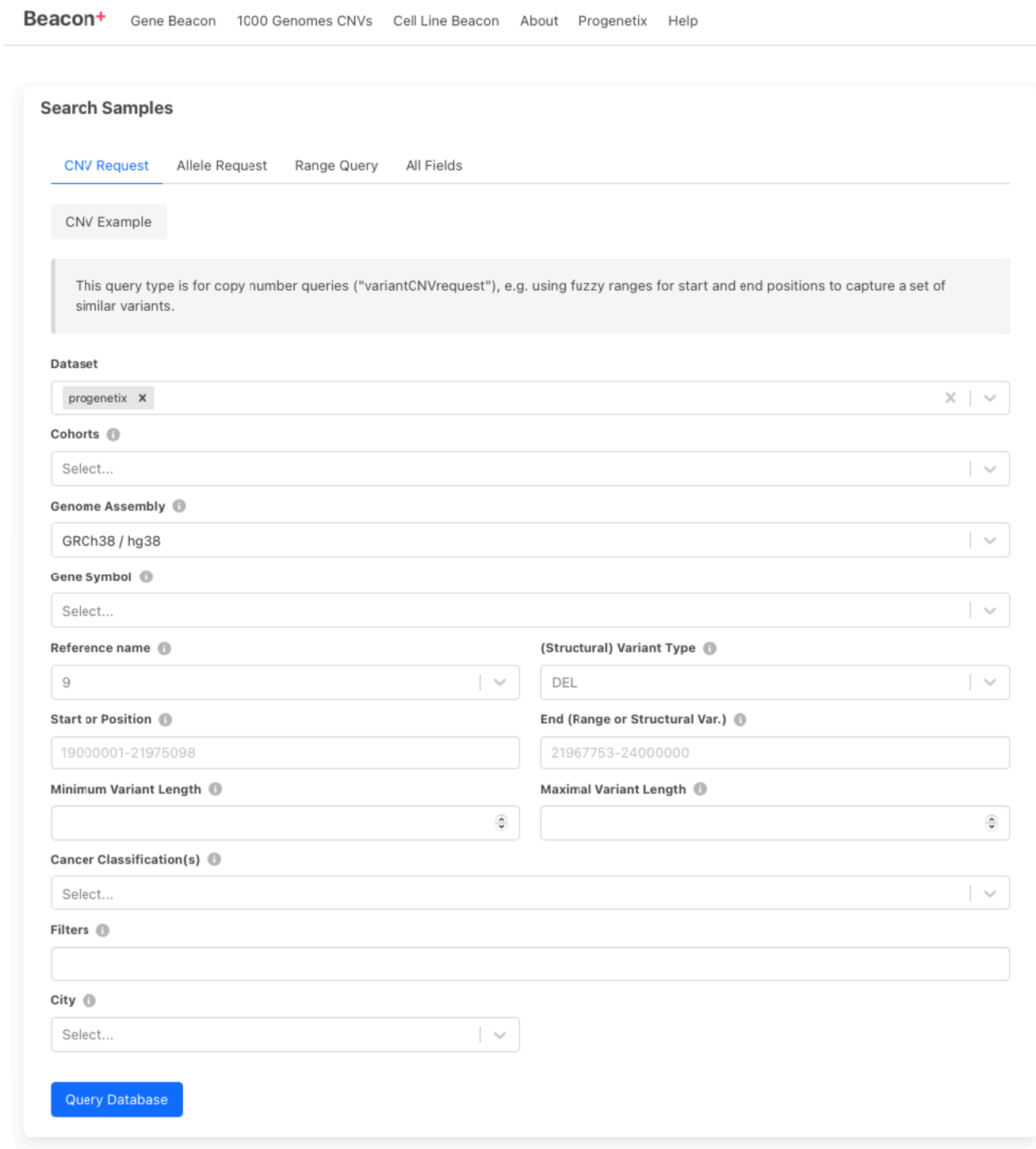
Beacon(+) “Metadata”

Beacon+ by Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

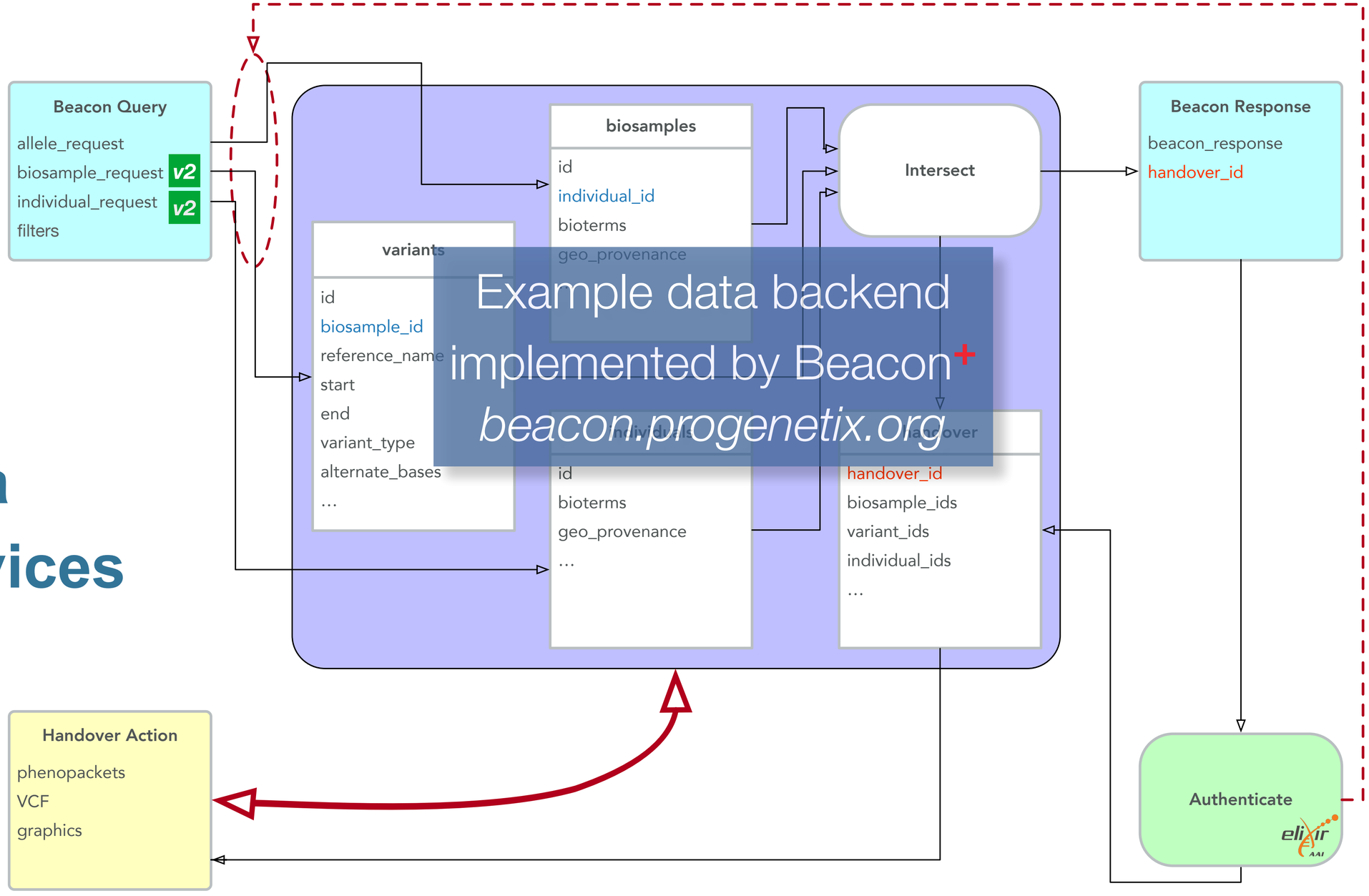
- Since 2016 the Progenetix resource has been used to model options for **Beacon development**
  - 138334 individual samples from 698 cancer types
- The consistent use of hierarchical diagnostic codes allows the use of Beacon "**filters**" for histopathological/clinically scoped queries
- Beacon's **handover** protocols can be utilized for data retrieval and, well, handing over to additional services, e.g.
  - downloads
  - visualization
  - use of external services (UCSC browser display...)

progenetix



Beacon & Handover

Beacons v1.1 supports **data delivery services**



Beacon I/O

Handover

Authentication

Michael Baudis

Global Alliance for Genomics & Health

SIB

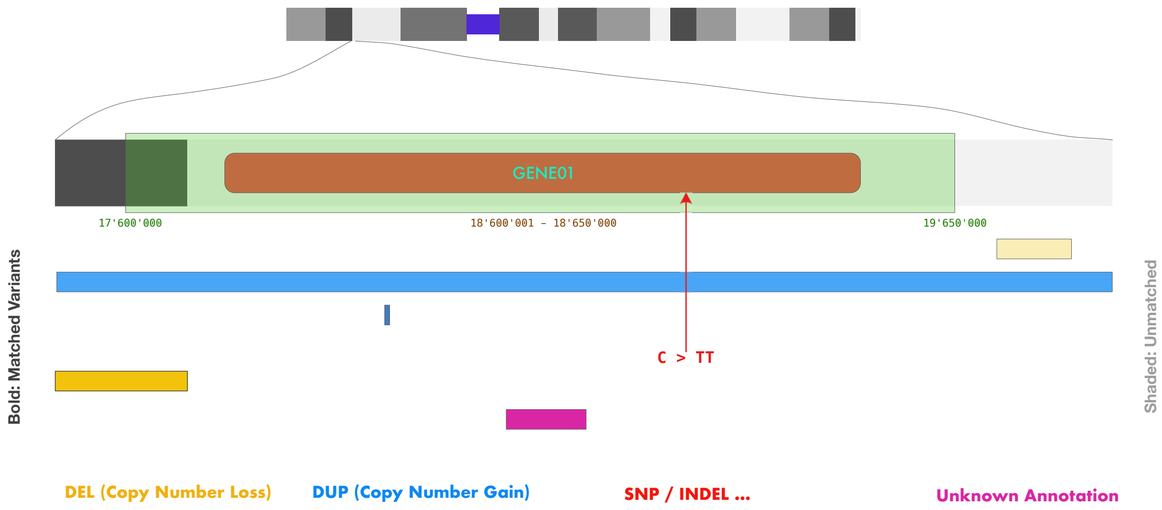
University of Zurich UZH

elixir

Beacon v2: Extended Variant Queries

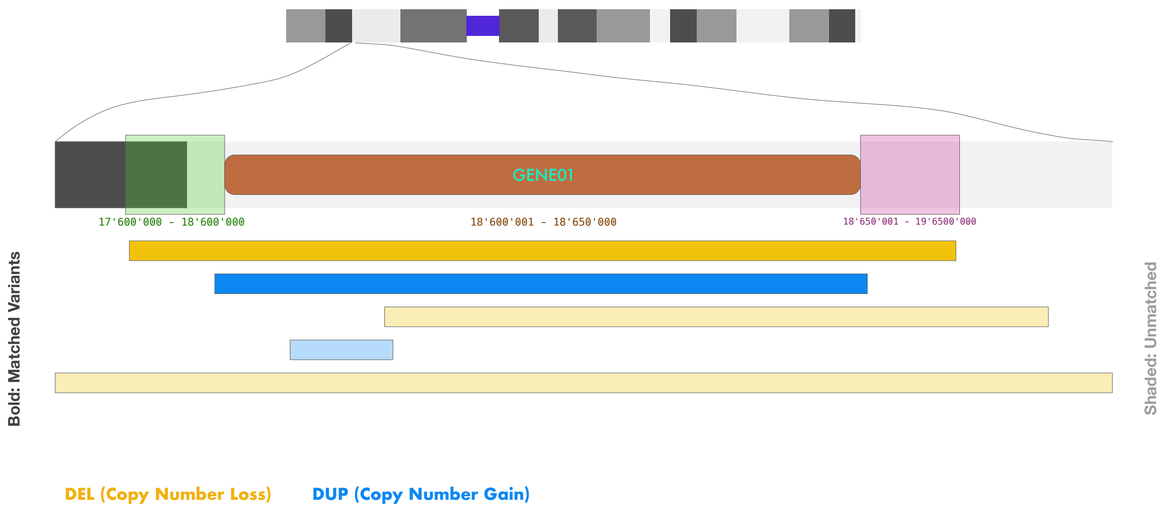
Range and Bracket queries enable positional wildcards and fuzziness

Genome Range Query
(matching variants in a region)



- Genome Range Queries provide a way to "fish" for variants overlapping an indicated region, e.g. the CDR of a gene of interest
- Additional parameters (e.g. variant type, reference or alternate bases) limit the scope of the responses
- new Beacon v2 size parameters to limit structural variants (e.g. "focal" CNVs)

Genome Bracket Query
(full match)



- Genome Bracket Queries allow to search for structural variants with start and end positions falling into defined sequence ranges
- allows to query any contiguous genomic variant (and in principle also can step in for range queries)
- typical use case is e.g the query for variants such as duplications covering the whole CDR of a gene, while limiting the allowed start or end regions

Onboarding

Demonstrating Compliance

- Progenetix Beacon+ has served as implementation driver since 2016
- Beacon v2 as service with protocol-driven registries for federation
- GA4GH approved Beacon v2 in April 2022

Global Alliance for Genomics & Health

Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon *v2* API

The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".

PaxDb

A Protein abundance reference resource

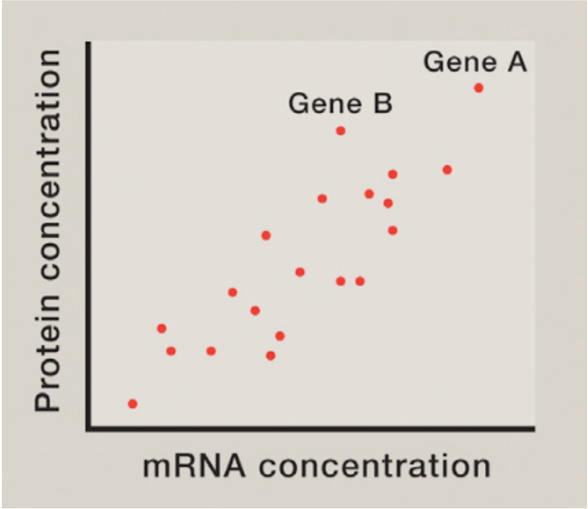
PaxDb

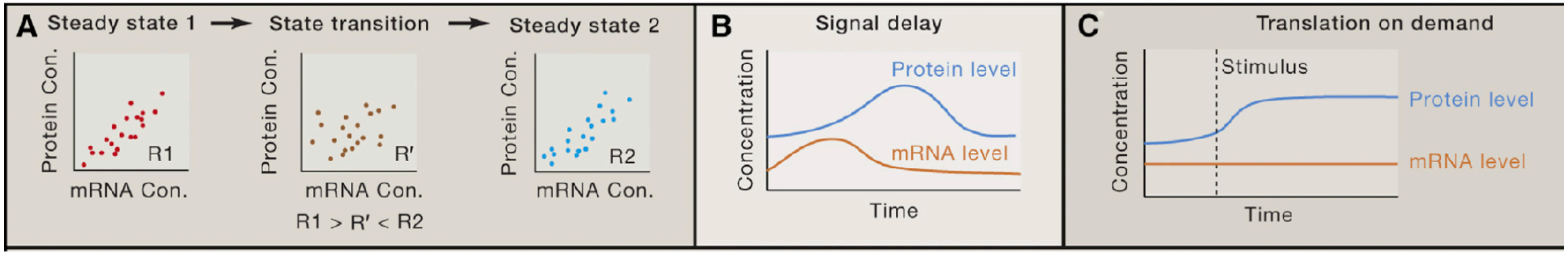
- Motivation for building a DB
  - Relevance
- What is the quantitative data and how is it represented?
  - Protein abundances
  - Orthology relationships
  - Techniques
- What is the metadata?
  - species, tissue, protein ID, ortholog
  - publication, experimental condition
- How to use the resources?
  - Web browsing, bulk download, upload own data

PaxDb

Motivation

- mRNA Levels primarily correlate with protein levels
- buffering of excess mRNA variation / noise
- regulated at functional level with modification, degradation etc.
- conserved in core processes
- conserved across species



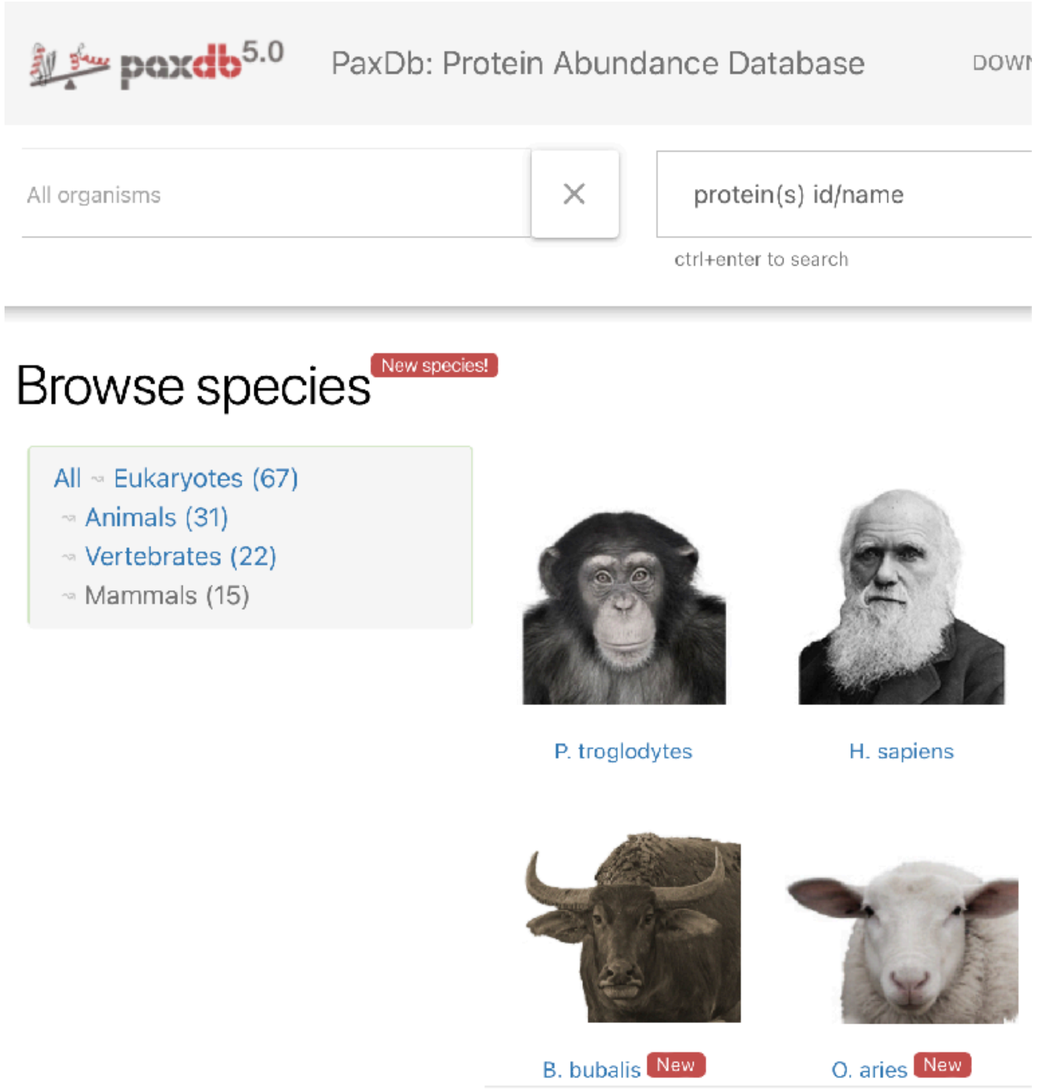


Liu et al. 2016 On the Dependency of Cellular Protein Levels on mRNA Abundance.Cell

PaxDb

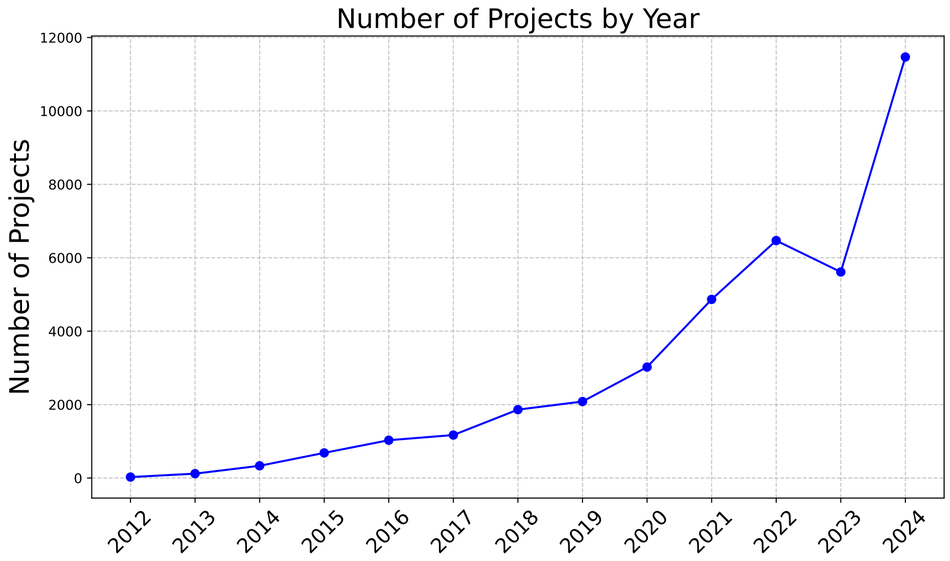
Motivation

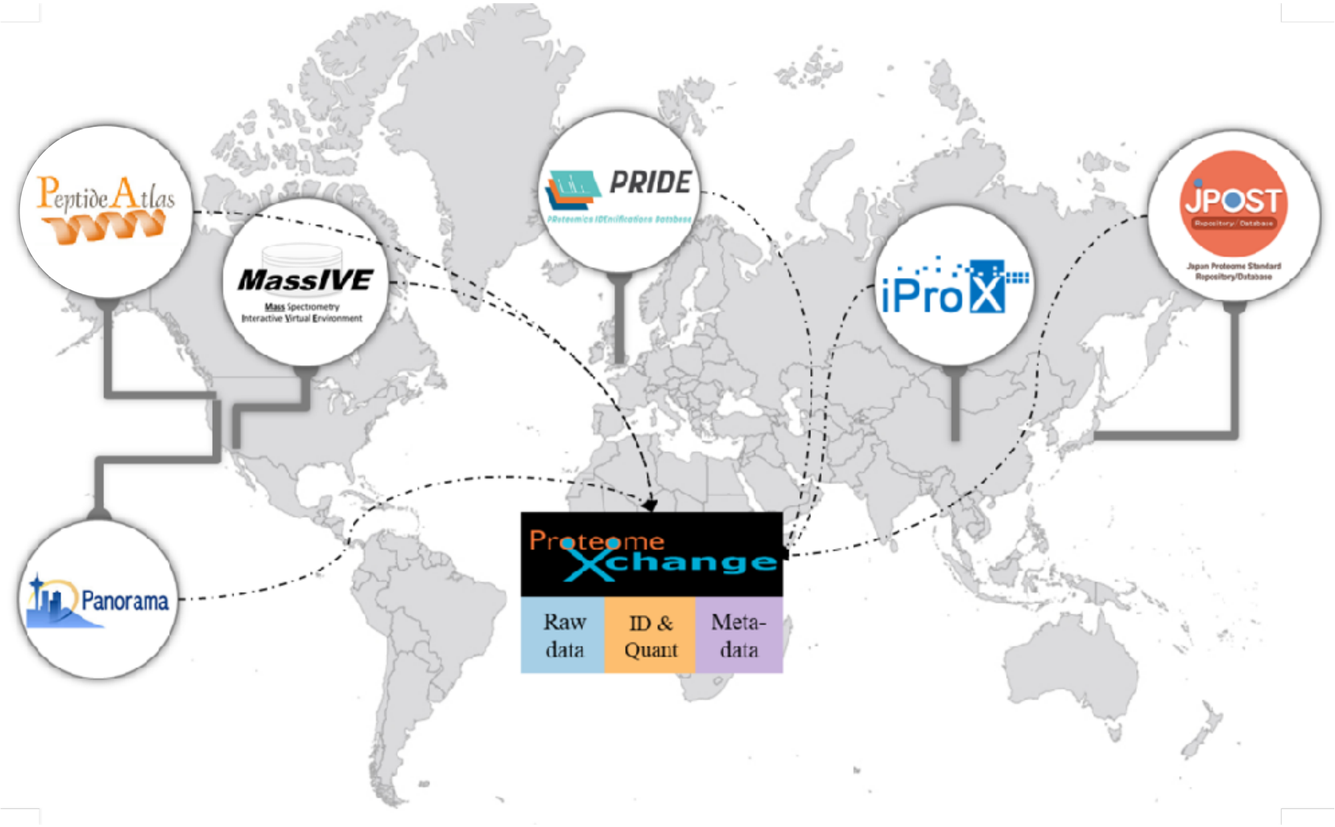
- Protein abundance across organisms
- Proteomics datasets are large and difficult to process and compare
- Reference for common and rare species
- Reference for cross-species comparison
- Integration on datasets of same type



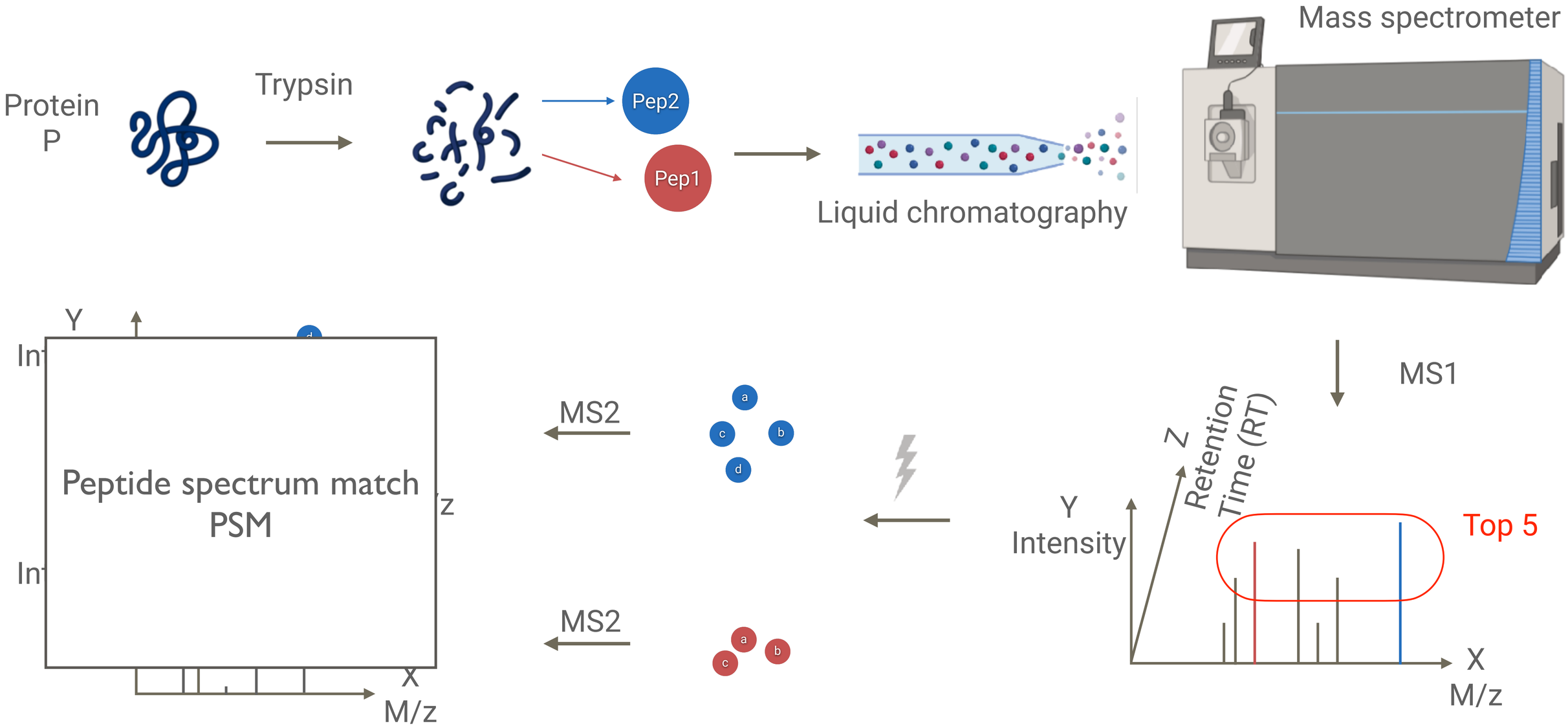
ProteomeXchange

- Data registry (consortium) of multi-regional data repositories
- At least raw data but also processed data



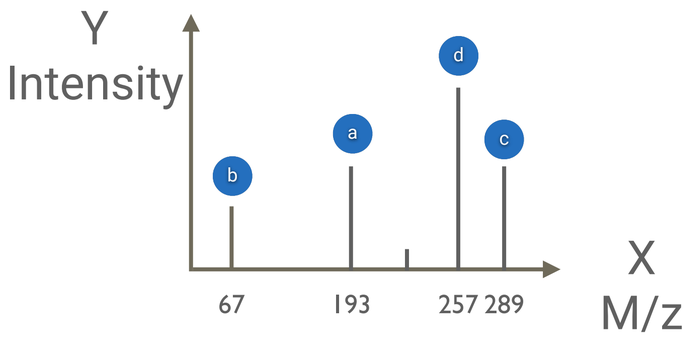


Quantitative Proteomics with LC-MS/MS



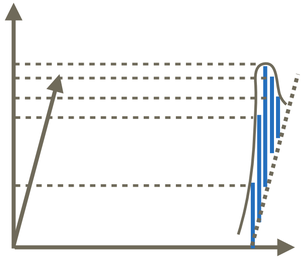
Data acquisition by Mass spectrometry

Peptide spectrum match (PSM)



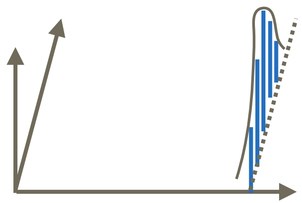
1. Peptide sequence from database search

2. Sum of MS1 Intensity



Intensity-based

2. count of peptide appearing as top peaks



count-based

- Sequence + intensity
- Sequence + count
- Protein ID + intensity
- Protein ID + count

Data collection

Manually downloaded

Filename	Description
nph17756-sup-0006-TableS5.xlsx Excel 2007 spreadsheet , 1.4 MB	**Table S5** Full data list of wheat grain protein turnover rates during grain development.

Annotation					
First ID	Protein gorup	Protein name	Intensity_R1	Intensity_R2	Intensity_R3
TraesCS1A01G00	TraesCS1A01G00	Nucleic acid-bind	341240	194450	426600
TraesCS1A01G00	TraesCS1A01G00	Paired amphipath	431440	393670	207190
TraesCS1D01G00	TraesCS1D01G00	Transcription init	57589	56370	29279
TraesCS1A01G00	TraesCS1A01G00	E3 ubiquitin-prot	1099200	1120300	1462500
TraesCS1A01G00	TraesCS1A01G00	Gamma-gliadin	670790	622010	945570
TraesCS1A01G00	TraesCS1A01G00	Peptidyl-prolyl ci	186530	167950	194550
TraesCS1A01G00	TraesCS1A01G00	Low molecular w	1524600	929110	1274100
TraesCS1A01G00	TraesCS1A01G00	MICOS complex s	231330	181940	222810
TraesCS1D01G00	TraesCS1D01G00	Ankyrin repeat fa	290850	421780	305910
TraesCS1A01G01	TraesCS1A01G01	Defensin	271140	263810	267260
TraesCS5D01G13	TraesCS5D01G13	V-type proton ATI	1726900	1476500	2310400

↓

Normalization

Downloaded in bulk from repos

Project Files

FTP GLOBUS

Name	Type	Size (M)	Download
VELOS16664.raw	RAW	451	FTP
VELOS16644.raw	RAW	449	FTP
VELOS16646.raw	RAW	455	FTP

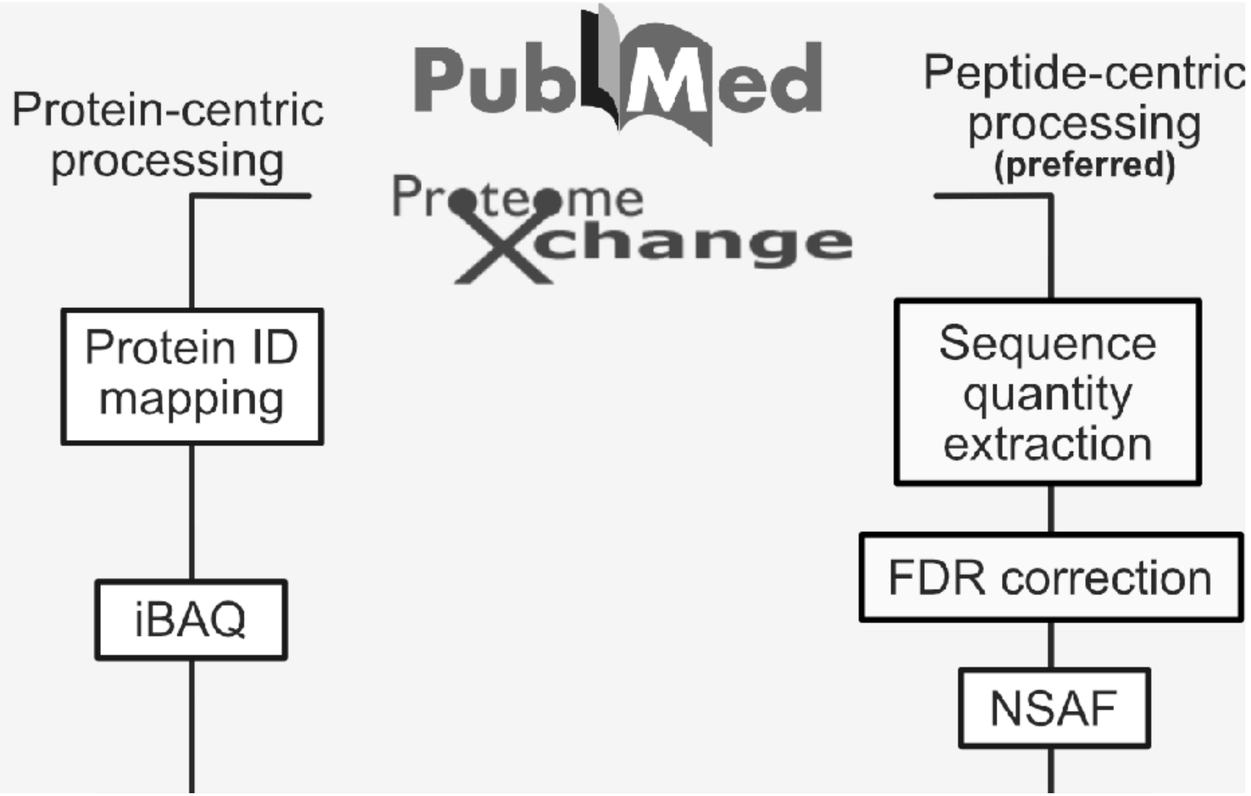
+

Name	Fraction	Experiment
VELOS16626		ZS1
VELOS16627		ZS1
VELOS16629		TiO1
VELOS16630		TiO1

↓

Mass spectrometry pipeline

Computation pipeline



Quality evaluation based on protein interaction

interacting proteins often have roughly similar abundances:

origin recognition complex

ORC1: 8.6 ppm	ORC4: 12.3 ppm
ORC2: 1.4 ppm	ORC5: 2.7 ppm
ORC3: 3.2 ppm	ORC6: 6.4 ppm

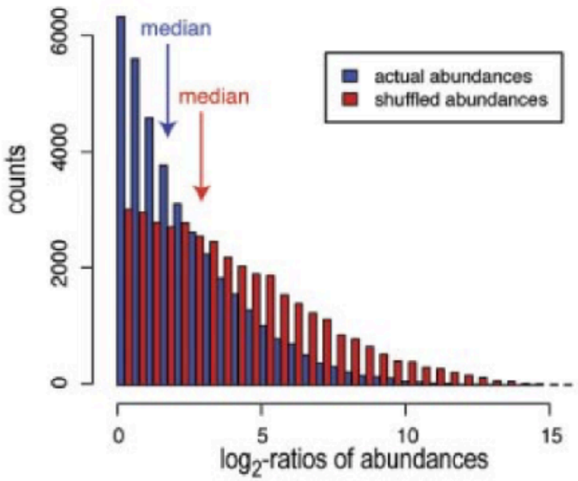
5.7 ± 4.2 ppm

replication factor A

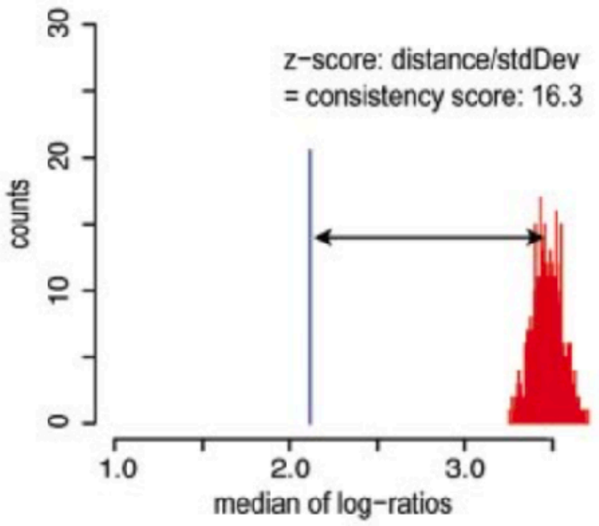
RFA1: 57 ppm
RFA2: 97 ppm
RFA3: 123 ppm

92 ± 33 ppm

pairwise comparisons of all interacting proteins in yeast

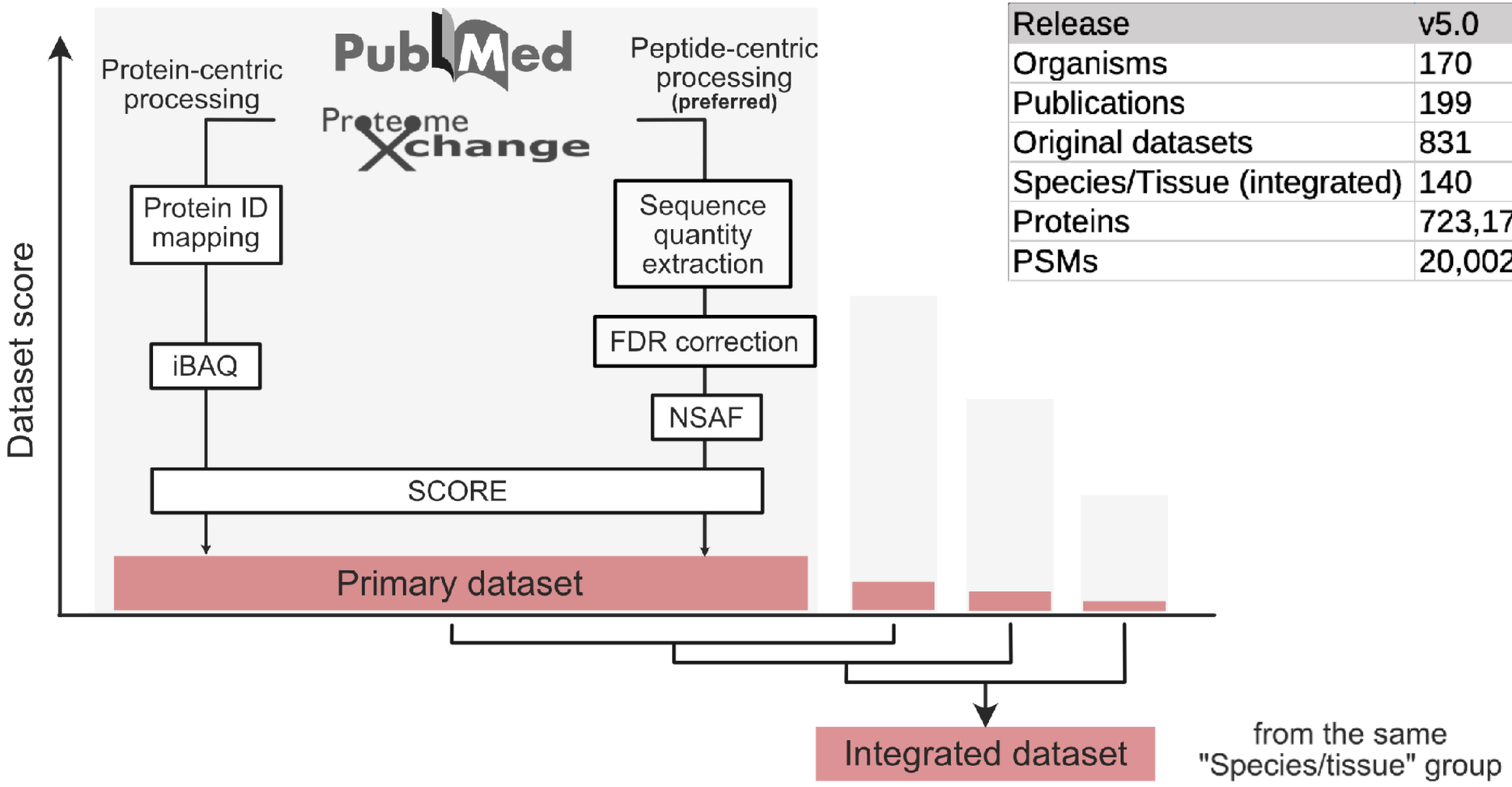


Consistency with protein interactions



Wang et al, MCP 2012

Data integration pipeline



Release	v5.0
Organisms	170
Publications	199
Original datasets	831
Species/Tissue (integrated)	140
Proteins	723,173
PSMs	20,002,648

Metadata

- Ontology
  - Species name → taxonomical ID
    - Homo sapiens → 9606
  - Tissue / organ → Uberon, Plant ontology …
    - THYROID_GLAND → UBERON:0002046
    - PERICARP → PO:0009084
  - Protein name → Protein ID
    - APOA2 → ENSP00000356969

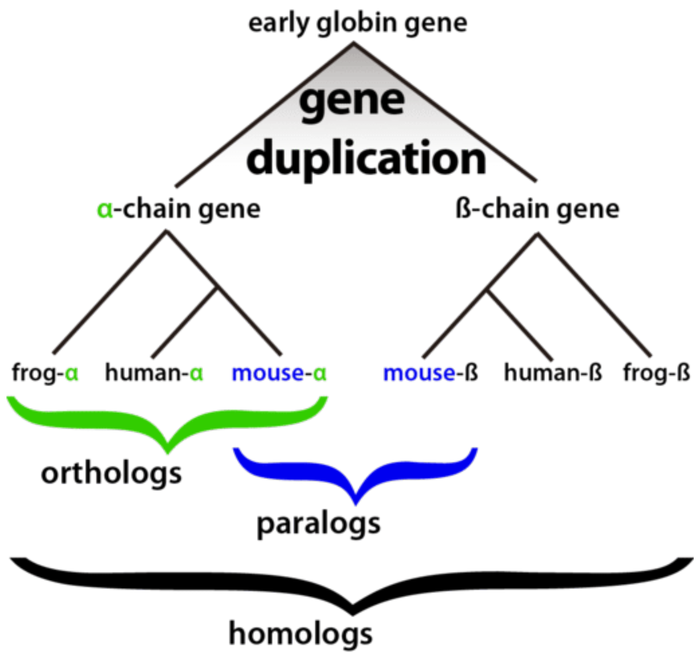
Metadata

- Taxonomical level
  - opisthokonta → 33154
  - primata → 9443
- eggNOG orthologs mapping
  - APOA2 → 9443.ENOG504MJTR (primate level)
  - APOA2 → 33208.ENOG503BTNZ (metazoa level)
- Publication
  - PaxDb 5.0… → PMID:37659604, 2023, Mol Cell Proteomics
- Experimental condition (free text)
  - Spectral counting, SILAC, DIA

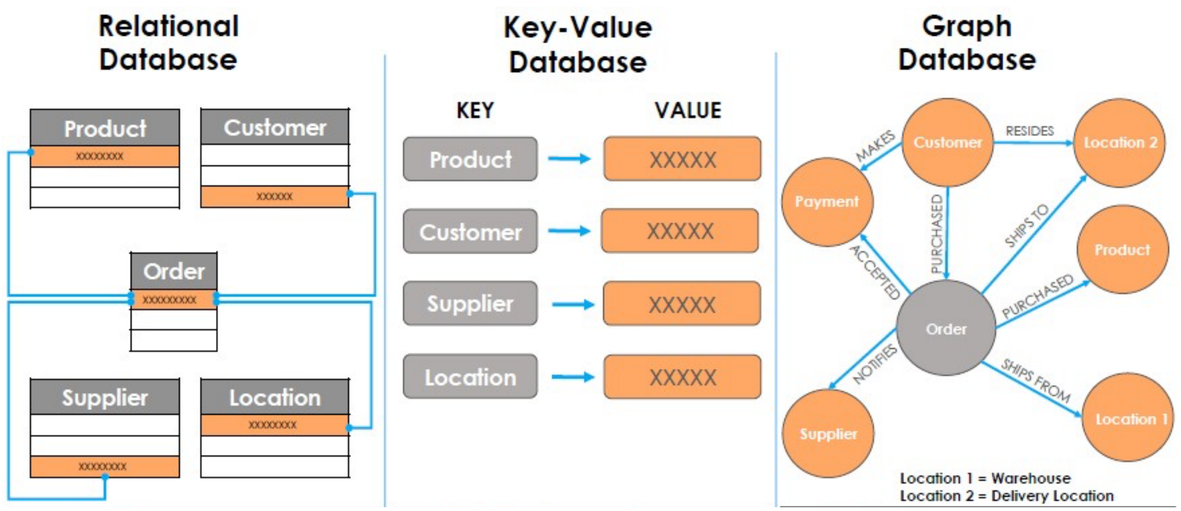
https://commons.wikimedia.org/wiki/File:Homology.png

Ortholog relation

- Genes in different species that evolved from a common ancestral gene by speciation
- Orthologs typically retain the same function
- At higher taxonomic level, the clusters of orthologous genes (Cogs) are larger and orthologs are more distant.
- In PaxDb, all orthologs are mapped to all levels up to the last universal common ancestor (Luca).



Data stored in a Graph database neo4j



e.g. SQL Redis neo4j

Search “friends of friends”…

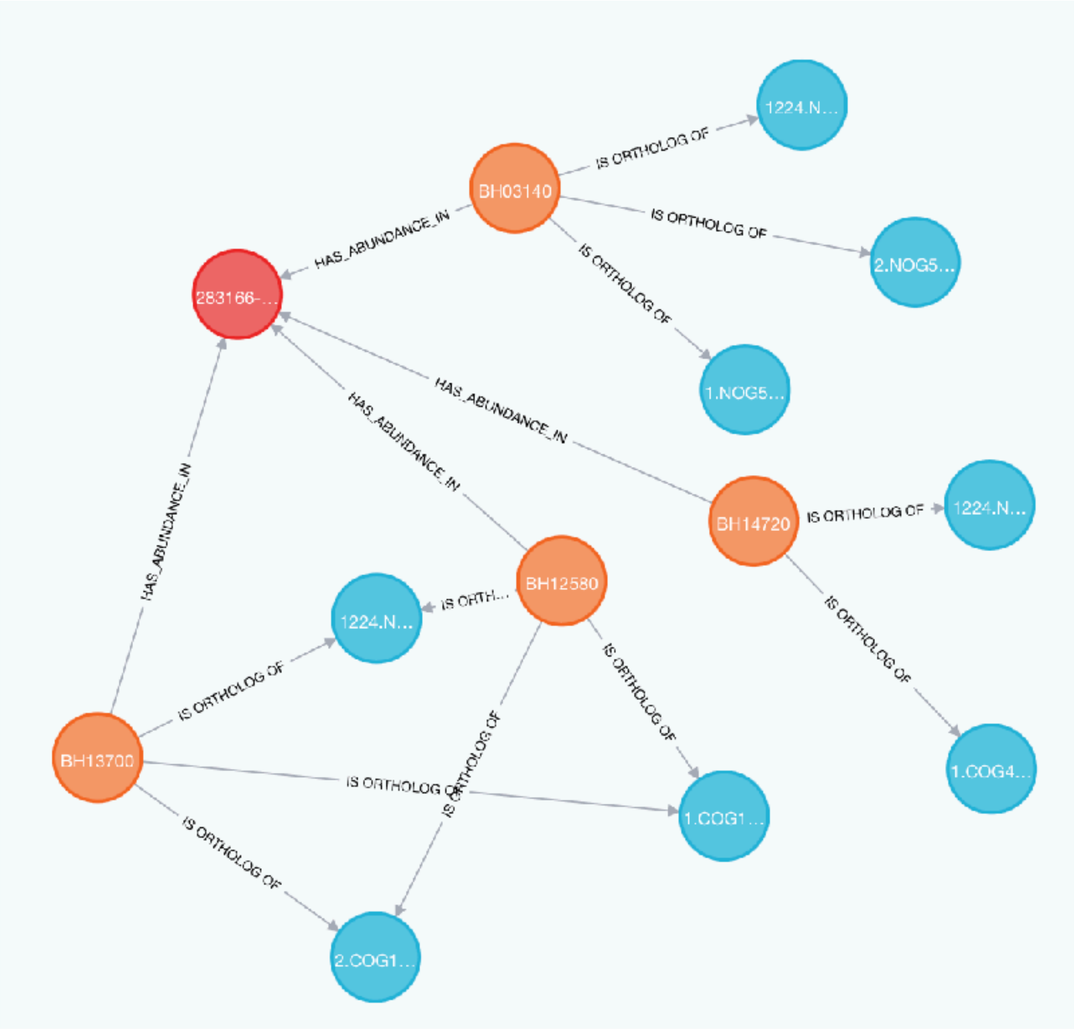
Depth	Execution Time – MySQL	Execution Time –Neo4j	Faster by…
2	0.016	0.010	60%
3	30.267	0.168	180x
4	1,543.505	1.359	1134x
5	Not Finished in 1 Hour	2.132	

Jonas Partner and Aleksa Vukotic. Neo4j in Action, 2014

Performant (comparison with SQL)

https://www.nextplatform.com/2018/09/19/the-graph-database-poised-to-pounce-on-the-mainstream/

Database structure



Protein(4) NOG(22) Dataset(1)

HAS_ABUNDANCE_IN(4) IS ORTHOLOG OF(11)

NOG **<id>:** 2254040 **level:** BACTERIA **levelId:** 2 **name:** 2.NOG52678

Protein **<id>:** 1087193 **eid:** 283166.BH14720 **iid:** 12091534 **name:** BH14720

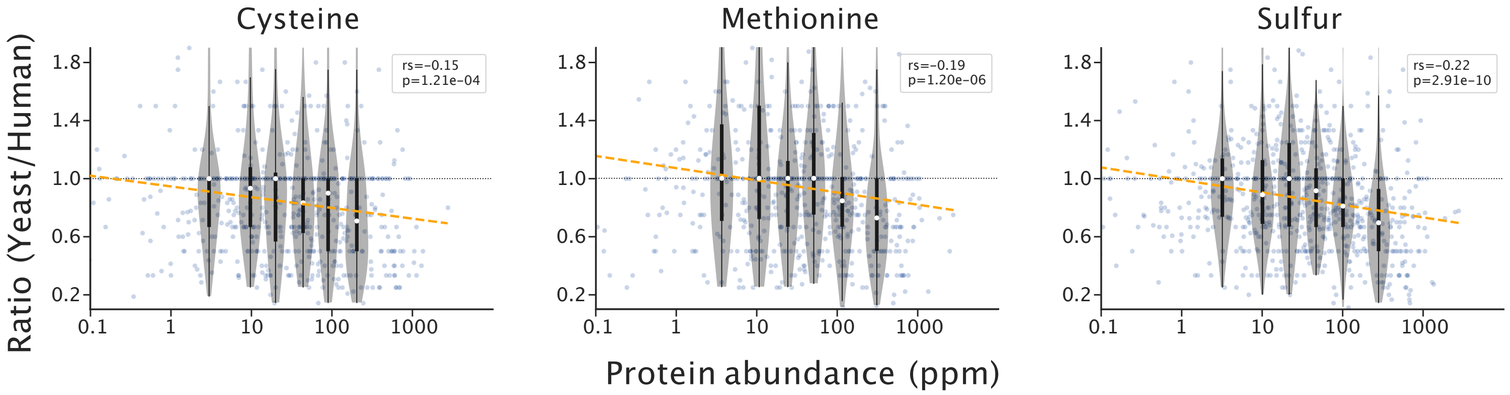
Dataset **<id>:** 2624041 **coverage:** 86 **filename:** 283166-Bhenselae_Albrethsen_2013 **iid:** 534428800 **integrated:** false **organ:** WHOLE_ORGANISM **score:** 8.9

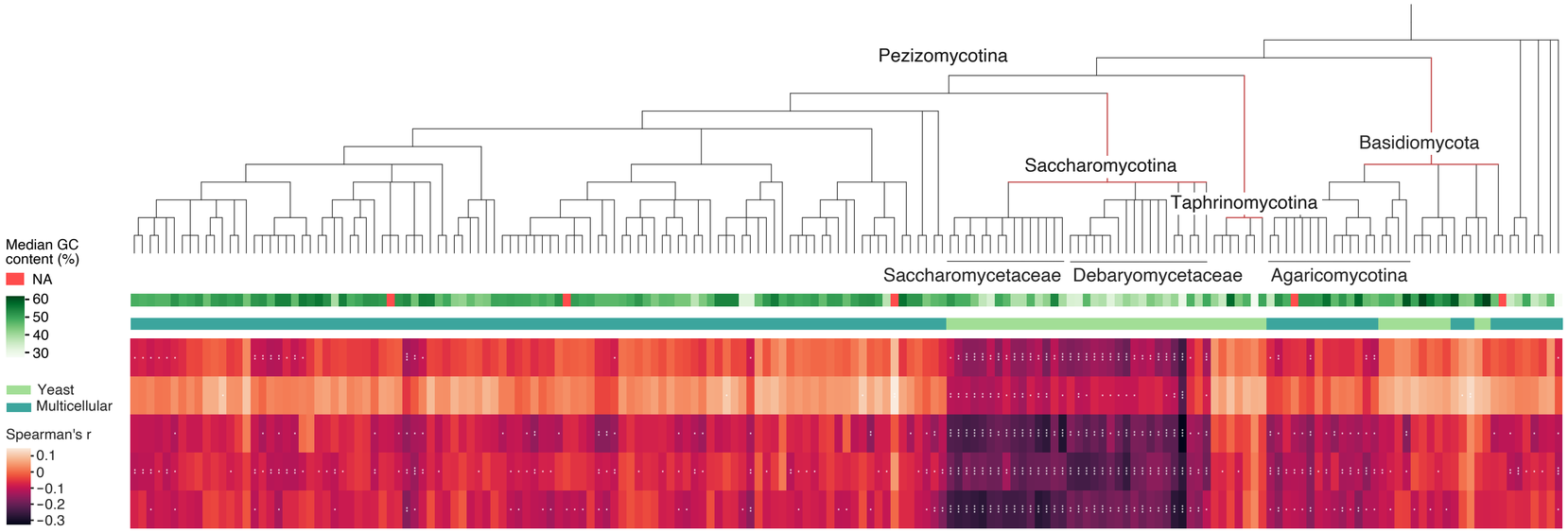
HAS_ABUNDANCE_IN **<id>:** 6874388 **ppm:** 590.0 **rank:** 306/1275

2,625,001 nodes and 11,824,389 edges

Data Use Cases

Protein abundance data reveals evolutionary signal





Other uses

Model protein turnover and half life

Predict codon bias

Predict binding affinity

Reference for stoichiometry

Verify own proteomics experiments

Access the data

Bulk download

- https://pax-db.org/downloads/

Index of /downloads/latest/

../		
datasets/	23-Jul-2023 06:53	-
paxdb-mapped_peptides-v5.0/	14-Feb-2023 15:00	-
paxdb-orthologs-v5.0/	09-Feb-2023 22:11	-
paxdb-protein-sequences-v5.0/	14-Feb-2023 13:53	-
paxdb-uniprot-links-v5.0/	14-Feb-2023 13:57	-
paxdb-mapped_peptides-v5.0.zip	14-Feb-2023 15:04	201795666
paxdb-orthologs-v5.0.zip	14-Feb-2023 13:59	25308621
paxdb-protein-sequences-v5.0.zip	14-Feb-2023 14:04	522039256
paxdb-uniprot-crossreferences.txt	03-Jun-2024 08:04	54403288
paxdb-uniprot-links-v5.0.zip	14-Feb-2023 13:58	10860175

Access the data

Individual dataset download by filtering

Datasets

heart

Name	Tissue type	Interaction consistency score	Coverage	Download
H.sapiens - Heart (Integrated)	Heart	33.6	68%	Download
H.sapiens - Heart, Fetal, SC (Kim,nature,2014)	Heart	26.7	51%	Download
H.sapiens - Heart, SC (Wangetal,molsystbiol2019)	Heart	17.1	47%	Download
H.sapiens - Heart, SC (Peptideatlas,aug,2014)	Heart	24.2	40%	Download
H.sapiens - Heart, SC (Kim,nature,2014)	Heart	30.5	33%	Download
H.sapiens - Heart, SC (Aye,mol_bio_syst,2010)	Heart	14.2	17%	Download
H.sapiens - Heart, SC (Kline,j.proteome_res,2008)	Heart	9.3	17%	Download
H.sapiens - Heart, normalized data APEX (Aye,mol_bio_syst,2010)	Heart	13	11%	Download

Microservice APIs

Ortholog API

For human protein FABP1,

- which taxonomy levels does it map to?
- what orthologs does it have at primate level?
- In what tissues does at least one ortholog have abundance values at primate level?
- what are the primate-level orthologs' abundances in the liver?

- https://orthologs-api.pax-db.org/protein/9606.ENSP00000295834/ortholog_groups/
- https://orthologs-api.pax-db.org/protein/9606.ENSP00000295834/ortholog_groups/PRIMATES/list_orthologs
- https://orthologs-api.pax-db.org/protein/9606.ENSP00000295834/ortholog_groups/PRIMATES/list_tissues
- https://orthologs-api.pax-db.org/protein/9606.ENSP00000295834/ortholog_groups/PRIMATES/LIVER

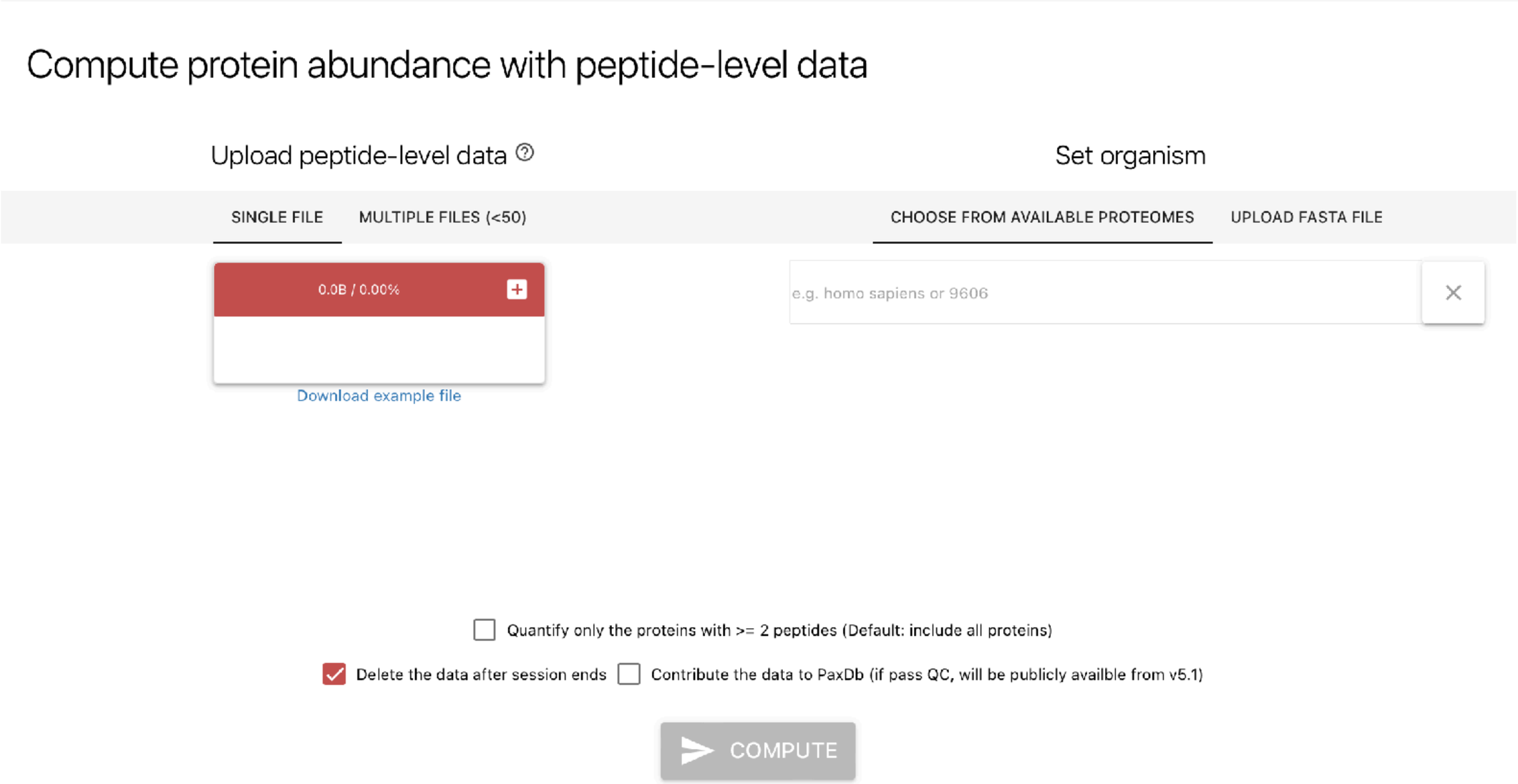
Microservice APIs

Data API

1. Show info of all datasets of *Arabidopsis thaliana*.
2. What are all the information about dataset xxx?
3. What are all protein abundance and annotation in the dataset xxx?
4. How is the protein abundance distrubution of dataset xxx?
5. Where does the protein xxx stand in the distribution?
6. What are all abundances of the protein by string ID xxx?
7. What are all abundances of the protein by Ubiprot ID xxx?
8. What are abundances of multiple proteins x, y, z … across all datasets?

1. https://api.pax-db.org/species/3702
2. https://api.pax-db.org/dataset/9606/986013392/abundances
3. https://api.pax-db.org/dataset/9606/986013392/
4. https://api.pax-db.org/dataset/986013392/histogram/
5. https://api.pax-db.org/dataset/986013392/histogram/?highlightProteinId=ENSP000003700
6. https://api.pax-db.org/protein/string/9606.ENSP00000295897
7. https://api.pax-db.org/protein/uniprot/Q851P9_ORYSJ
8. https://api.pax-db.org/proteins?ids=9606.ENSP00000269305,9606.ENSP00000258149

Upload own data



Upload own data - result

Single dataset

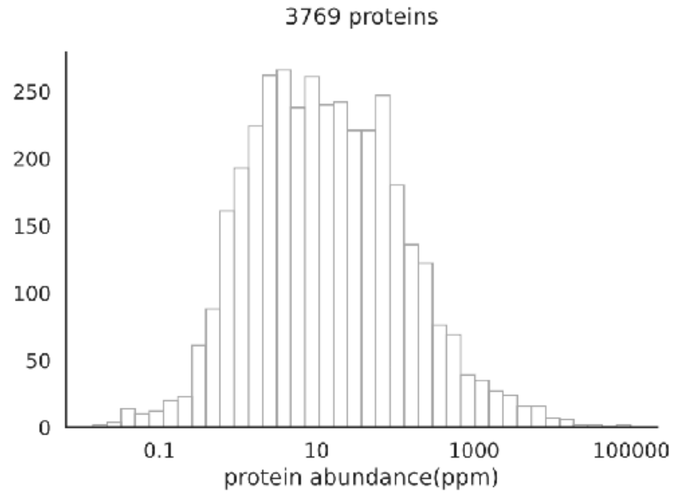
Results and summary

The uploaded dataset has a score of 10.35.

For details of interaction score, please refer to Wang, M. et al. Mol Cell Proteomics 2012

Top 20 Most abundant proteins

#	Protein	ppm
1	ALB	97267
2	PTGDS	69907
3	CST3	33347
4	TTR	31420
5	TMSB4X	22646



Multiple datasets

Summary of computed files

File name	Status	Interaction score	Range (ppm)	No. Proteins	Top 3 proteins
human_example	success	10.35	0.01 - 97267	3769	ALB, PTGDS, CST3
human_example2	success	8.68	0.03 - 22784	3769	GM2A, LOR, NAXE

Components of an Online Bioinformatics Resource

Going Full Stack?

Components of an Online Bioinformatics Resource

A Stack to work with/through

- dedicated server or cloud storage
- own domain | institutional sub-domain or fixed address | cloud service sub-domain
  - ➡ progenetix.org, pax-db.org I mls.uzh.ch/en/research/baudis I baudisgroup.github.io
- webserver gateway for server-side generated, active content delivery
  - ➡ Perl CGI, Python, PHP …
- Web front-end with html+css or dynamic site with javascript frameworks
- database
  - ➡ SQL databases such as PostGres, MySQL
  - ➡ document databases such as MongoDB, CouchDB, Elastic search...
  - ➡ graph database such as neo4j

(Bio)informatics Skill Set

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

GitHub

text mining

Perl

regular expressions

s/knowledge/mastery/

React

mongoDB

M↓

APACHE HTTP SERVER PROJECT

R

MkDocs

Project documentation with Markdown.

python™

HTML CSS JS

array & sequencing pipelines

(Bio)informatics Skill Set

What has been needed to develop & maintain pax-db.org?

GitHub

text mining

Flask
web development,
one drop at a time

docker

QUASAR

node
express

APACHE
HTTP SERVER PROJECT

Java

neo4j

python™

HTML CSS JS

Mass spectrometry pipelines

Last but NOT Least…

Documentation is, actually, rather important

Documentation Strategies

(Not so) Best Practices

```
f_d = f_d_s[c_t]
r = {}
for k in res_schema.keys():
    if k in f_d:
        r.update({k:f_d[k]})
```

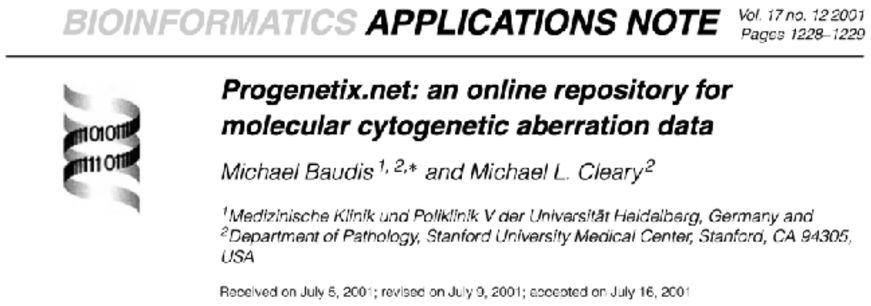
- What is documentation? I'll remember this! ¯_(ツ)_/¯

mbaudis@netscape.net

- Just email me if help is needed, unexpectedly
- We had money for a chat bot.

```
normalize_variant_values_for_export(v, byc, drop_fields=None):
```

- Clean code documents itself - Just use explicit variable/function names.
- Clean code documents itself - Never use explicit variable/function names.
- Perl POD it is. There is a command to show the notes in your terminal…
- I wrote a paper about the resource. In 2001.
- Haven't you found the GoogleGroups account?
- Documentation? StackOverflow, whelp!



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

Documentation Strategies

Currently en Vogue

- Cloud-based documentation systems with online compilation
- written in simplified markup languages
  - ➡ Markdown (Yeah!)
  - ➡ Restructured Text (Meeh...)
- local and/or service based compilation and hosting
- build systems & output hosting
  - ➡ ReadTheDocs
    - direct building from .rst document tree or MkDocs based
  - ➡ Github Pages
    - direct using Jekyll or over MkDocs through GH actions

GitHub

Read the Docs
Create, host, and browse documentation.

MkDocs
Project documentation with Markdown.

jekyll

Documentation Strategies

Read the Docs
Create, host, and browse documentation.

Sign up

or Log in

Technical documentation lives here

Read the Docs simplifies software documentation by automating building, versioning, and hosting of your docs for you.

Free docs hosting for open source

We will host your documentation for free, forever. There are no tricks. We help over 100,000 open source projects share their docs, including a custom domain and theme.

Always up to date

Whenever you push code to your favorite version control service, whether that is GitHub, BitBucket, or GitLab, we will automatically build your docs so your code and documentation are never out of sync.

Downloadable formats

We build and host your docs for the web, but they are also viewable as PDFs, as single page HTML, and for eReaders. No additional configuration is required.

Multiple versions

We can host and build multiple versions of your docs so having a 1.0 version of your docs and a 2.0 version of your docs is as easy as having a separate branch or tag in your version control system.

Example: GA4GH Variation Representation Standard ->

GA4GH Variation Representation Specification

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

stable

Search docs

Introduction
Terminology & Information Model
Schema
Implementation Guide
Releases
Appendices

Read the Docs v: stable

Docs » GA4GH Variation Representation Specification

Edit on GitHub

GA4GH Variation Representation Specification

The Variation Representation Specification (VRS, pronounced "verse") is a standard developed by the Global Alliance for Genomic Health to facilitate and improve sharing of genetic information. The Specification consists of a JSON Schema for representing many classes of genetic variation, conventions to maximize the utility of the schema, and a Python implementation that promotes adoption of the standard.

Citation

The GA4GH Variation Representation Specification (VRS): a computational framework for variation representation and federated identification. Wagner AH, Babb L, Alterovitz G, Baudis M, Brush M, Cameron DL, ..., Hart RK. *Cell Genomics*. Volume 1 (2021). doi:10.1016/j.xgen.2021.100027

Output

ahwagner add docs ··· ✓ on Jan 29 History

..

_static	Use shared metaschema tooling (#354)	13 months ago
appendices	remove reference to develop branch (#344)	14 months ago
images	Closes #324: Removed Abundance from current schema; re-implemente...	14 months ago
impl-guide	fix link to Data Proxy class	14 months ago
releases	Closes #320: Add note about attributes that permit identifiable and n...	17 months ago
conf.py	Closes #345: Fix sphinx theming (#346)	14 months ago
defs	Use shared metaschema tooling (#354)	13 months ago
index.rst	update citation	
introduction.rst	update doc urls to use vrs.ga4gh.org	2 years ago

Source

MkDocs & Material for MkDocs & Github Actions

FOLDERS

- progenetix-web
  - .github
  - .next
  - docs
    - css
    - img
    - javascripts
    - news
    - beaconplus.md
    - changelog.md
    - classifications-an
    - CNAME
    - index.md
    - progenetix-data-r
    - progenetix-websi
    - publication-collec
    - services.md
    - technical-notes.m
    - ui.md
    - use-cases.md
  - extra
  - node_modules
  - out
  - public
  - src
  - .babelrc
  - .env.development
  - .env.production
  - .eslintrc.json
  - .gitignore
  - .prettierrc
  - jest.config.js
  - mkdocs.yaml
  - next.config.js
  - package-lock.json
  - package.json
  - README.md

mkdocs.yaml

```yaml
site_name: Progenetix Documentation
site_description: 'Documentation for the Progenetix oncogen
site_author: Michael Baudis
copyright: '&copy; Copyright 2022, Michael Baudis and proge
repo_name: 'progenetix-web'
repo_url: https://github.com/progenetix/progenetix-web

##########################################################

nav:
  - Documentation Home: index.md
  - News & Changes: news
  - Pages & Forms: ui

  - Publication Collection: publication-collection
  - Data Review: progenetix-data-review
  - Technical Notes: technical-notes
  - Progenetix Website Builds: progenetix-website-builds
  - Progenetix Data &#8599;: http://progenetix.org
  - Baudisgroup @ UZH &#8599;: http://info.baudisgroup.org

##########################################################

markdown_extensions:
  - toc:
      toc_depth: 2-3
      permalink: true
  - admonition
  - attr_list
  - footnotes
  - md_in_html
  - pymdownx.critic
  - pymdownx.caret
  - pymdownx.details
  - pymdownx.keys
  - pymdownx.magiclink:
      hide_protocol: true
  - pymdownx.mark
  - pymdownx.tilde
  - pymdownx.saneheaders
```

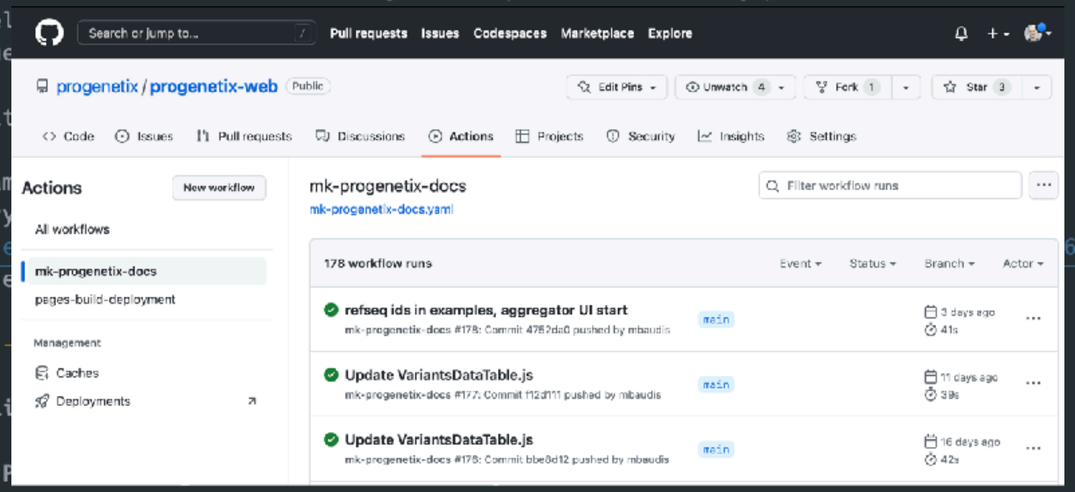
classifications-and-ontologies.md

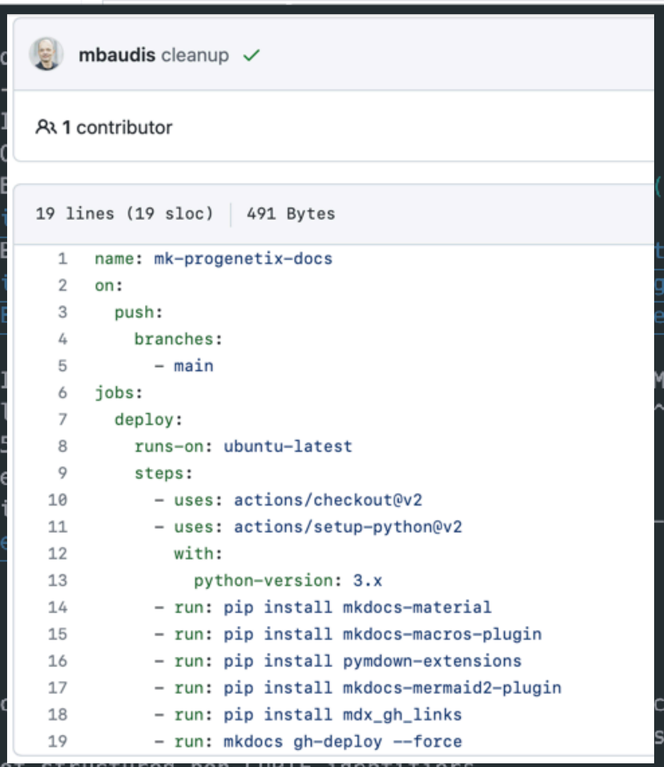
```md
# Classifications, Ontologies and Standards

The Progenetix resource utilizes standardized diagnostic coding systems, with a
move towards hierarchical ontologies. As part of the coding process we have
developed and provide several code mapping resources through repositories, the
Progenetix website and APIs.

Additionally to diagnostic and other clinical concepts, Progenetix increasingly
uses hierarchical terms and concepts for the annotation and querying of technical
parameters such as platform technologies. Overall, the Progenetix resource uses a
query syntax based around the [Beacon v2 "filters"](https://beacon-project.io/v2/
filters.html) concept with a [CURIE](https://www.w3.org/TR/2010/NOTE-curie-20101216/)
-based syntax

### Public Ontologies with CURIE-based syntax

| CURIE prefix | Code/Ontology                          | Examples |
| ------------ | ------------------------------------- | -------- |
| NCIT         | NCIt Neoplasm[^1]                      | NCIT:C27676 |
| HP           | HPO[^2]                                | HP:0012209 |
| PMID         | NCBI Pubmed ID                         | [PMID:18810378](http://
progenetix.org/services/ids/PMID:18810378) |
| geo          | NCBI Gene Expression Omnibus[^3] | [geo:GPL6801](http://
progenetix.org/services/ids/geo:GPL6801), [geo:GSE19399](http://progenetix.org/
services/ids/geo:GSE19399), [geo:GSM491153](http://progenetix.org/services/ids/geo:
GSM491153) |
| arrayexpress | EBI ArrayExpress[^4]                   | arrayexpress:E-MEXP-1008 |
| cellosaurus  | Cellosaurus - a knowledge resource on cell lines [^5]|
cellosaurus:CVCL_1650 |
| UBERON       | Uberon Anatomical Ontology[^6]         |UBERON:0000992|
| cbioportal   | cBioPortal[^9]                         | [cbioportal:msk_impact_2017](http
://progenetix.org/services/ids/cbioportal:msk_impact_2017) |

### Private filters

Since some classifications cannot directly be referenced, and in accordance with
the upcoming Beacon v2 concept of "private filters", Progenetix uses
additionally a set of structured non-CURIE identifiers.
```

Local Testing

```
→  progenetix-web git:(main) mkdocs serve
INFO     -  Building documentation...
INFO     -  [macros] - Macros arguments: {'module_name': 'main',
            'modules': [], 'include_dir': '', 'include_yaml': [],
            'j2_block_start_string': '', 'j2_block_end_string': '',
            'j2_variable_start_string': '', 'j2_variable_end_string': '',
            'on_undefined': 'keep', 'on_error_fail': False, 'verbose': False}
INFO     -  [macros] - Extra variables (config file):
            ['excerpt_separator', 'blog_list_length', 'social']
INFO     -  [macros] - Extra filters (module): ['pretty']
INFO     -  MERMAID2  - Initialization arguments: {}
INFO     -  MERMAID2  - Using javascript library (8.8.0):
               https://unpkg.com/mermaid@8.8.0/dist/mermaid.min.js
INFO     -  Cleaning site directory
INFO     -  The following pages exist in the docs directory, but are not
            included in the "nav" configuration:
              - beaconplus.md
              - changelog.md
              - classifications-and-ontologies.md
              - progenetix-data-review.md
              - progenetix-website-builds.md
              - publication-collection.md
INFO     -  MERMAID2  - Found superfences config: {'custom_fences': [{'name':
            'mermaid', 'class': 'mermaid', 'format': <function fence_mermaid at
            0x104075ab0>}]}
INFO     -  MERMAID2  - Page 'Technical Notes': found 2 diagrams, adding scripts
INFO     -  Documentation built in 0.83 seconds
INFO     -  [09:05:32] Watching paths for changes: 'docs', 'mkdocs.yaml'
INFO     -  [09:05:32] Serving on http://127.0.0.1:8000/
INFO     -  [09:05:33] Browser connected:
            http://127.0.0.1:8000/classifications-and-ontologies/
```

Web Deployment (Github)






```yaml
name: mk-progenetix-docs
on:
  push:
    branches:
      - main
jobs:
  deploy:
    runs-on: ubuntu-latest
    steps:
      - uses: actions/checkout@v2
      - uses: actions/setup-python@v2
        with:
          python-version: 3.x
      - run: pip install mkdocs-material
      - run: pip install mkdocs-macros-plugin
      - run: pip install pymdown-extensions
      - run: pip install mkdocs-mermaid2-plugin
      - run: pip install mdx_gh_links
      - run: mkdocs gh-deploy --force
```

Classifications, Ontologies and Standards

The Progenetix resource utilizes standardized diagnostic coding systems, with a move towards hierarchical ontologies. As part of the coding process we have developed and provide several code mapping resources through repositories, the Progenetix website and APIs.

Additionally to diagnostic and other clinical concepts, Progenetix increasingly uses hierarchical terms and concepts for the annotation and querying of technical parameters such as platform technologies. Overall, the Progenetix resource uses a query syntax based around the Beacon v2 "filters" concept with a CURIE based syntax.

List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

CURIE prefix	Code/Ontology	Examples
NCIT	NCIt Neoplasm[1]	NCIT:C27676

Table of contents

Documentation Strategies

Best Practices

- start early
- update often
- sometimes try to follow your own guide
- balance between inline documentation & doc system
- use Markdown
- plan for contingencies
  - ➡ cloud providers disappear | cancel services | change terms

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

https://en.wikipedia.org/wiki/List_of_search_engines#Defunct_or_acquired_search_engines