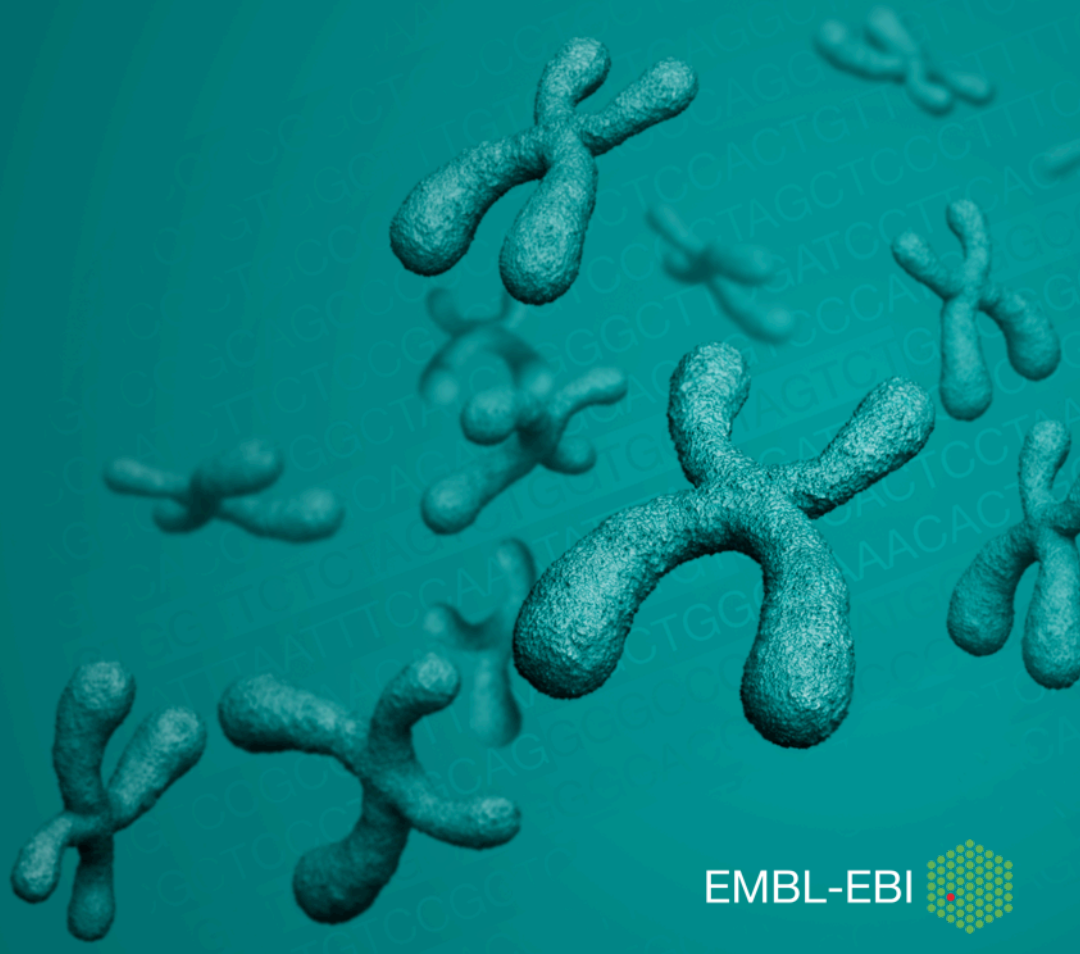
Ontologising the GWAS Catalog
‘A picture paints a thousand traits’

Helen Parkinson, EBI

17 July 2013



EMBL-EBI

Overview

- Introduction
- Infrastructure and Ontology
- GWAS diagram
- Outlook

The NHGRI GWAS catalog

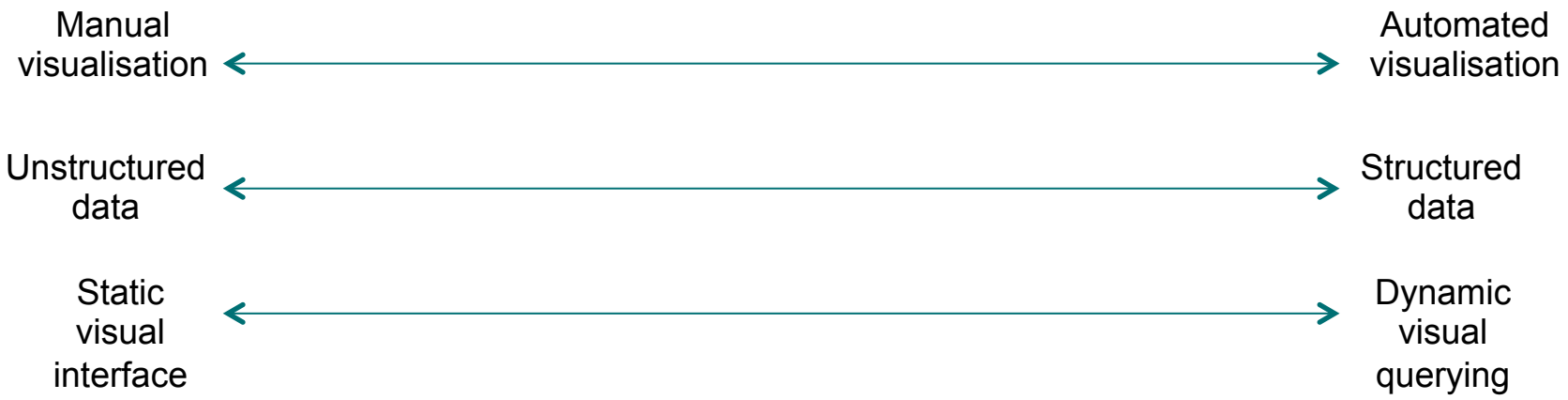
- Manual curation of published GWAS studies
  - Weekly literature search to identify new studies
  - Manual data extraction into web interface
  - Data entry double-checked by 2nd-level curator
- Quarterly release of GWAS diagrams
- Process failing to scale

release	Dec 2012
papers	1724
#SNPs p<5E-8	5035
#SNP-trait assocations p<5E-8	12593

Date Added to Catalog (since 11/25/08)	First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Reported Gene(s)	Mapped Gene(s)	Strongest SNP-Risk Allele	Context	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC]	CNV
11/09/11	Khor CC October 16, 2011 *Nat Genet* Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1.	Dengue shock syndrome	2,008 Vietnamese pediatric cases, 2,018 Vietnamese controls	1,737 Vietnamese cases, 2,934 Vietnamese controls	6p21.33 10q23.33 16p13.3 8q11.23	*MICB* *PLCE1* *NR* *NR*	MICB PLCE1 RBFOX1 SEC11B - RP1	rs3132468-? rs3765524-? rs6500818-? rs10104997-?	intron missense intron intergenic	0.13 0.70 NR NR	4×10^{-11} 3×10^{-10} 2×10^{-7} 9×10^{-7}	1.34 [1.23-1.46] 1.25 [1.16-1.33] 1.31 [NR] 1.2 [NR]	Illumina [481,342]	N

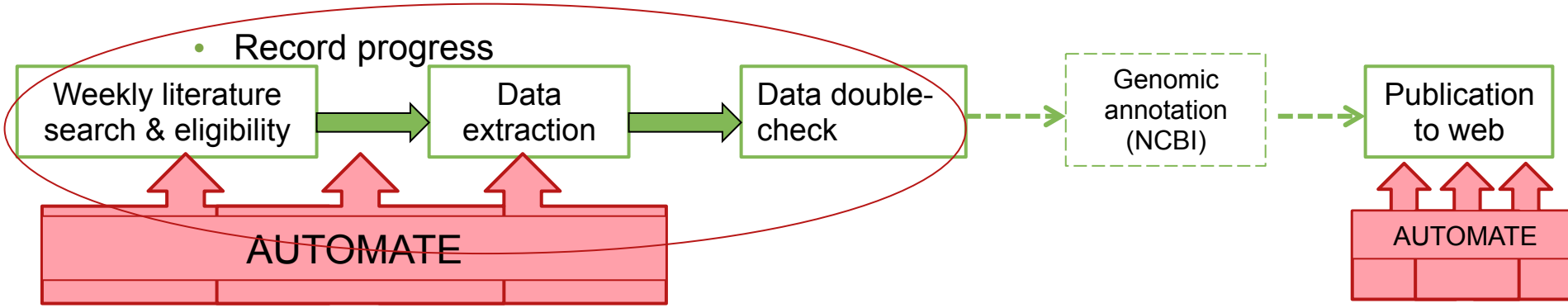
EBI/NHGRI collaboration

- 2-year collaboration between the GWAS catalog team at the NHGRI and the Functional Genomics Productions (development) and Vertebrate Genomics (curation & display through Ensembl variation) teams at EBI
- Aims



Curation infrastructure

- Development of tools to increase efficiency and accuracy of curation of data into the GWAS catalogue
  - Catalogue curation currently a labour intensive, entirely manual process
  - Development of an online tracking system to
    - Automatically perform Pubmed searches and enter papers into the system for review by curators
    - Triage papers
    - Assignment of papers to the appropriate curator for each stage of the curation process
    - Extract data from papers – SNP batchloader
    - Record progress



GWAS traits

- GWAS catalogue traits previously only available as an unstructured list

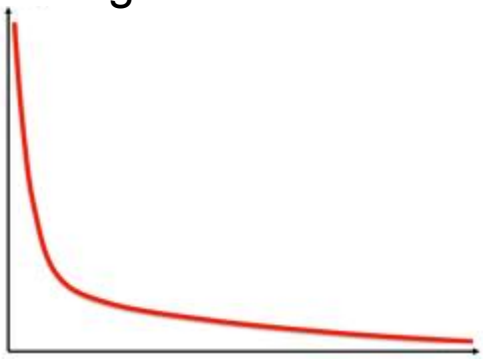
Dental caries
Depression and alcohol dependence
Depression--quantitative trait
Diabetes (gestational)
Diabetes (incident)
Diabetes related insulin traits
Diabetic nephropathy
Diabetic retinopathy
Dialysis-related mortality
Diastolic blood pressure
Digit length ratio
Dilated cardiomyopathy

Triglycerides-Blood Pressure (TG-BP)
Tuberculosis
Two-hour glucose challenge
Type 1 diabetes
Type 1 diabetes autoantibodies
Type 2 diabetes
Type 2 diabetes and 6 quantitative traits
Type 2 diabetes and gout
Type 2 diabetes and other traits
Ulcerative colitis

- Traits are highly diverse, including
  - Phenotypes, e.g. hair colour
  - Treatment responses, e.g. response to antineoplastic agents
  - Diseases, e.g. type 2 diabetes
  - Assays – glcyoslyated haemoglogin level
  - Chemical/drug names, e.g. C-reactive protein
- Traits are often compound and/or context-dependent

  e.g. “Type 2 diabetes and gout” or “Parkinson’s disease (interaction with caffeine)”

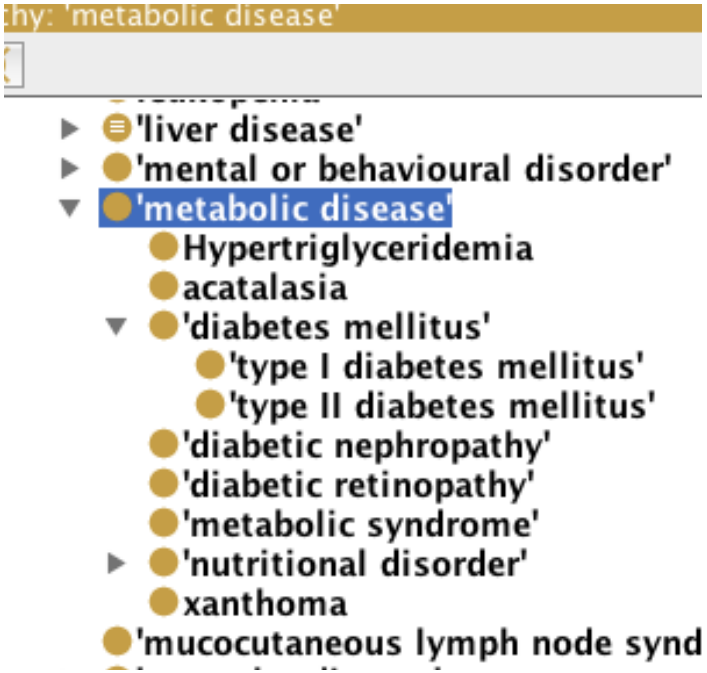
Long tail on the data



Ontology

- Integration of traits into the structured hierarchy of an ontology, with additional semantically meaningful links between traits allows much more complex and extensive querying, e.g.

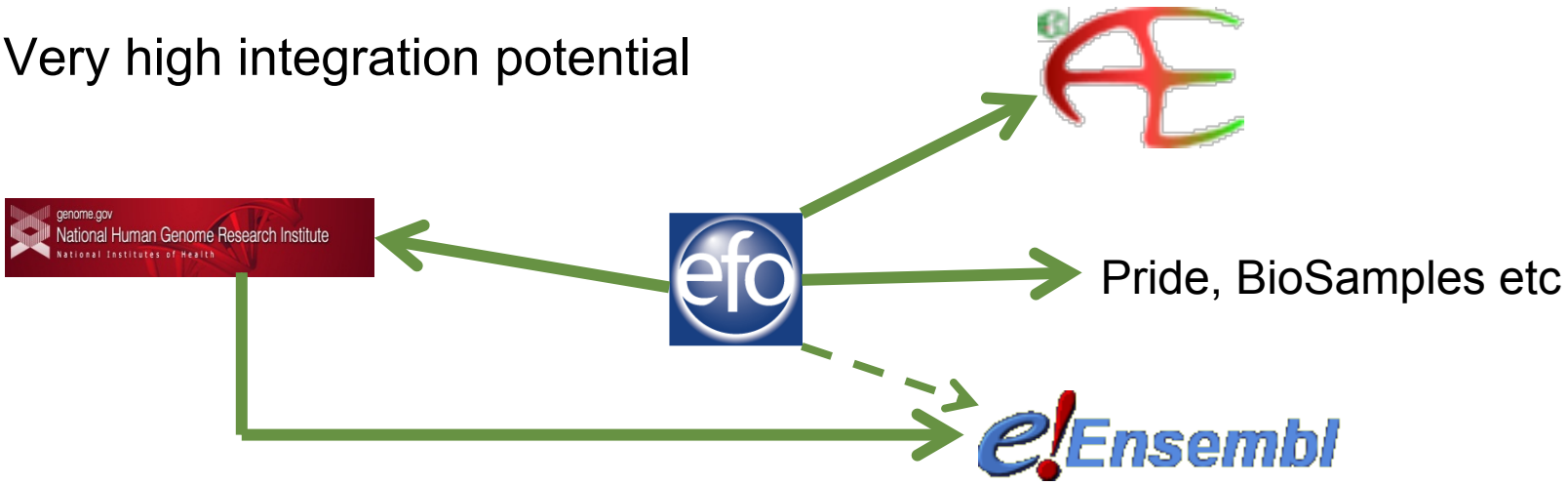
"Show me all SNPs associated with type 2 diabetes and metabolic syndrome"



- Two options for ontology integration
  - Create new "GWAS ontology"
  - Integrate with an existing ontology

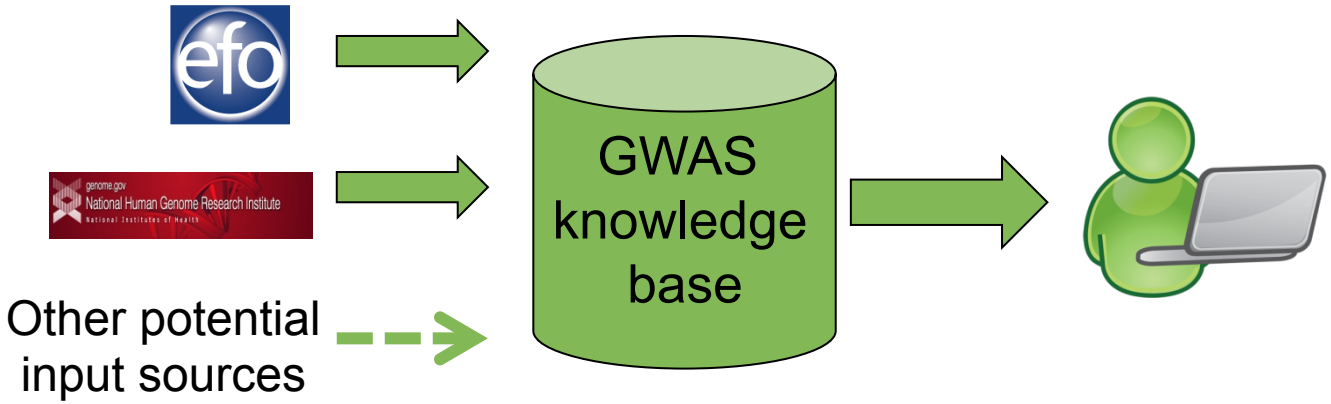
Integration with “Experimental Factor Ontology”

- EFO is actively developed
- Well-suited to covering diversity of GWAS traits
- 20% of GWAS traits already found in EFO prior to integration process
- ~500 new terms added over 5 releases = 100% coverage GWAS data
- Very high integration potential



New and more powerful queries

- Knowledge base that imports all the GWAS catalogue data and EFO



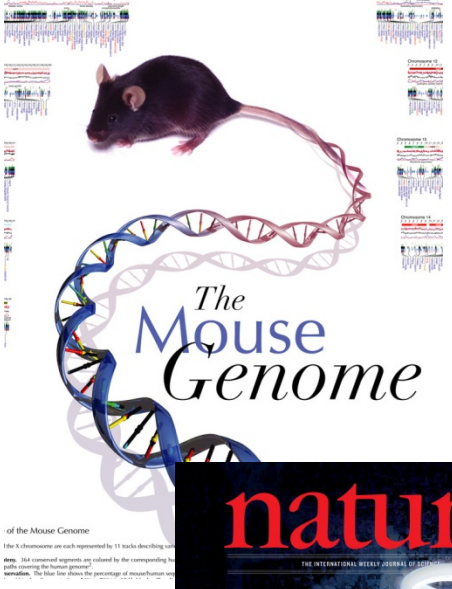
  - More powerful queries

    e.g. “Show me all SNPs associated with type 2 diabetes and metabolic syndrome, with a p-value of 10^{-5}, from papers published before January 2010”

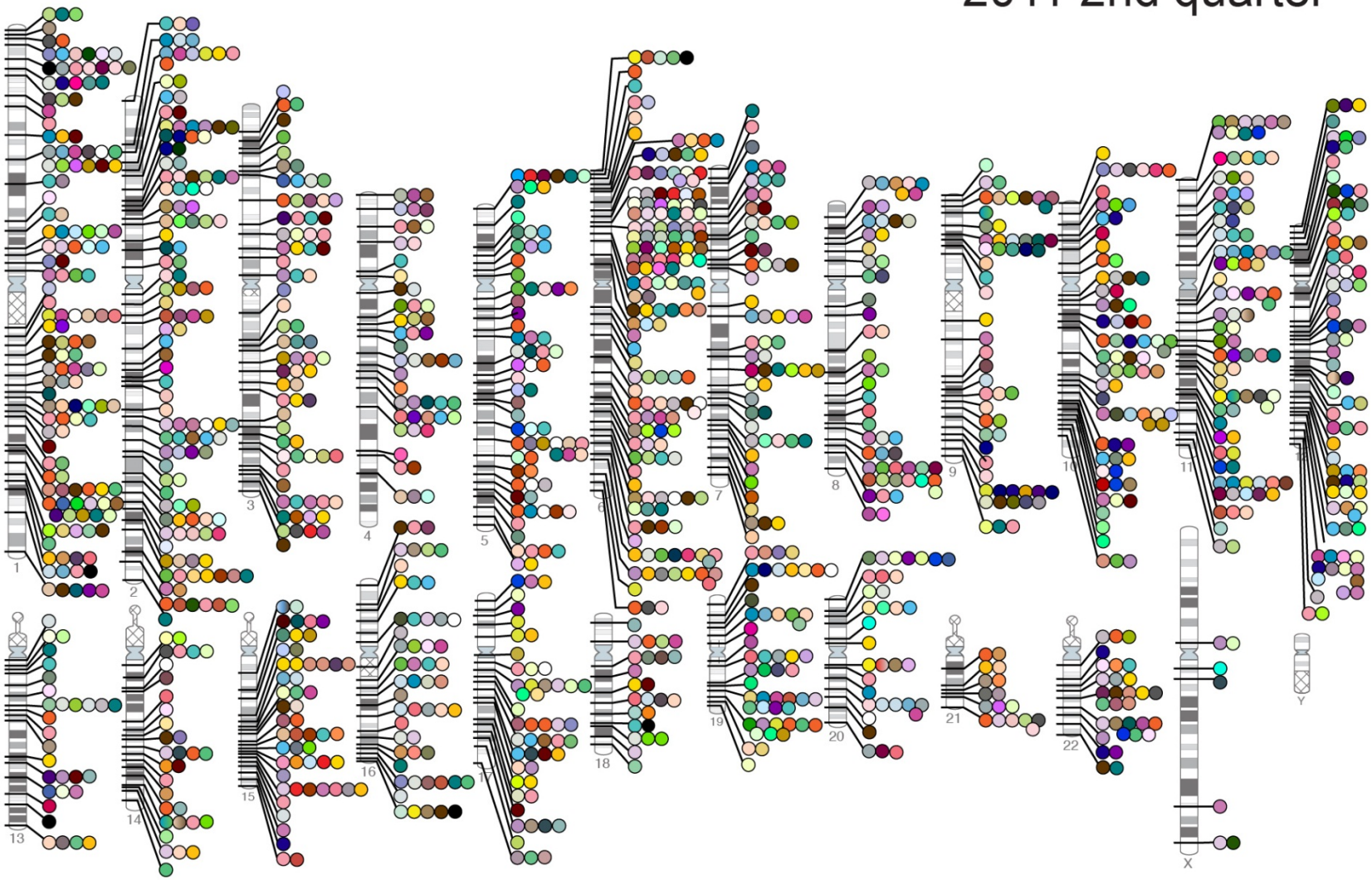
  - Facilitate visualisation
  - Increased integration potential, interoperability with other ontologies

GWAS diagram

- Visualisation of all SNP-trait associations with p-value < 10^{-8}
- Generated quarterly by a graphic artist following extensive manual curation of the data
- Static image in PDF or Powerpoint format
- Too many traits and colours to reliably identify any individual feature
- Great way of visualising the evolution of the catalogue over time



2011 2nd quarter

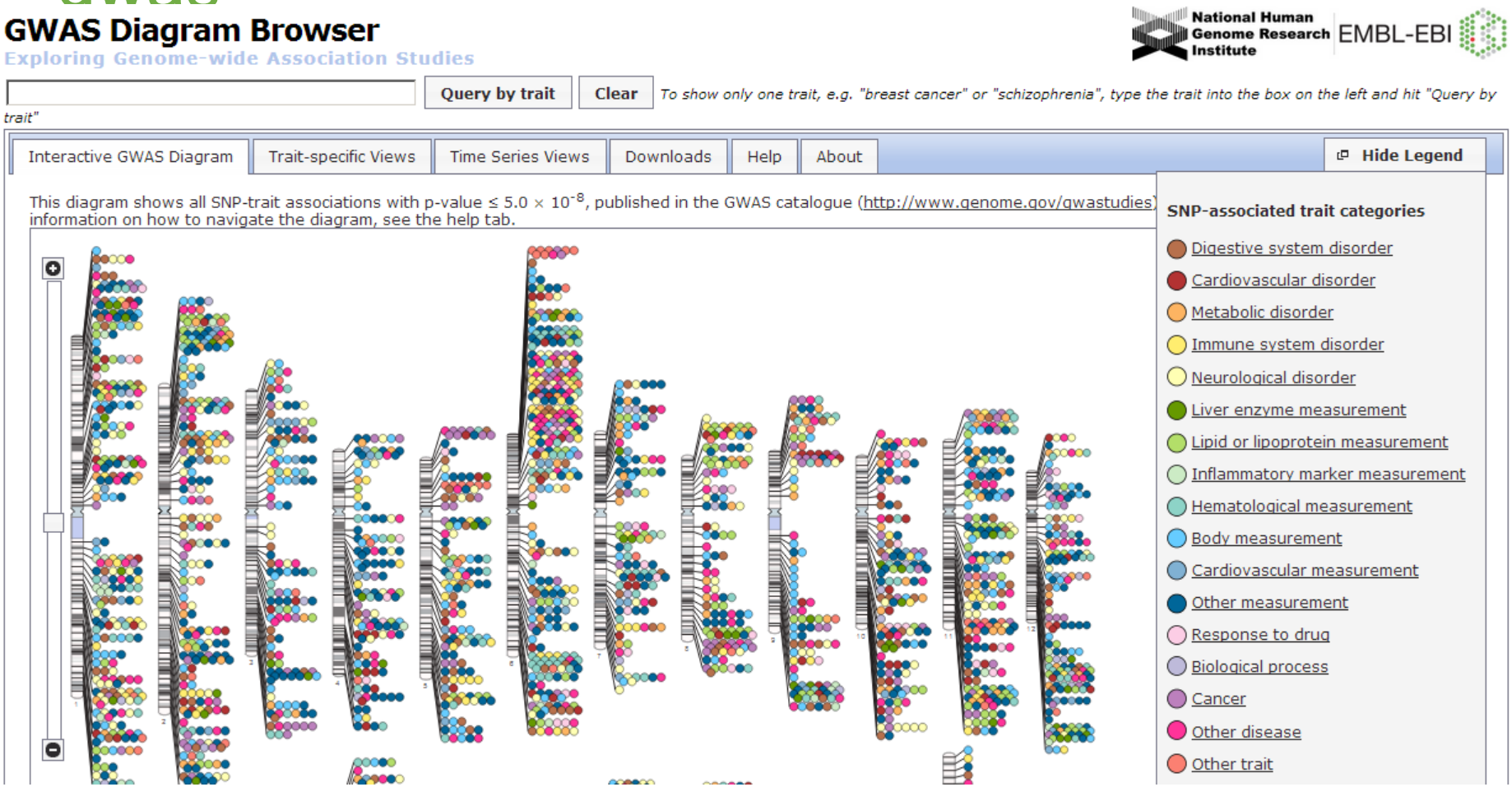



Abdominal aortic aneurysm
Acute lymphoblastic leukemia
Adhesion molecules
Adiponectin levels
Age-related macular degeneration
AIDS progression
Alcohol dependence
Alopecia areata
Alzheimer disease
Amyloid A levels
Amyotrophic lateral sclerosis
Angiotensin-converting enzyme activity
Ankylosing spondylitis
Arterial stiffness
Asparagus anosmia
Asthma
Atherosclerosis in HIV
Atrial fibrillation
Attention deficit hyperactivity disorder
Autism
Basal cell cancer
Behcet's disease
Bipolar disorder
Biliary atresia
Bilirubin
Bitter taste response
Birth weight
Bladder cancer
Bleomycin sensitivity
Blond or brown hair
Blood pressure
Blue or green eyes
BMI, waist circumference
Bone density
Breast cancer
C-reactive protein
Calcium levels
Cardiac structure/function
Cardiovascular risk factors
Carnitine levels
Carotenoid/tocopherol levels
Celiac disease
Celiac disease and rheumatoid arthritis
Cerebral atrophy measures
Chronic lymphocytic leukemia
Chronic myeloid leukemia
Cleft lip/palate
Coffee consumption
Cognitive function
Conduct disorder
Colorectal cancer
Corneal thickness
Coronary disease
Creutzfeldt-Jakob disease
Crohn's disease
Crohn's disease and celiac disease
Cutaneous nevi
Cystic fibrosis severity
Dermatitis
DHEA-s levels
Diabetic retinopathy
Dilated cardiomyopathy
Drug-induced liver injury
Drug-induced liver injury (amoxicillin-clavulanate)
Endometrial cancer
Endometriosis
Eosinophil count
Eosinophilic esophagitis
Erectile dysfunction and prostate cancer treatment
Erythrocyte parameters
Esophageal cancer
Essential tremor
Exfoliation glaucoma
Eye color traits
F cell distribution
Fibrinogen levels
Folate pathway vitamins
Follicular lymphoma
Fuch's corneal dystrophy
Freckles and burning
Gallstones
Gastric cancer
Glioma
Glycemic traits
Hair color
Hair morphology
Handedness in dyslexia
HDL cholesterol
Heart failure
Heart rate
Height
Hemostasis parameters
Hepatic steatosis
Hepatitis
Hepatocellular carcinoma
Hirschsprung's disease
HIV-1 control
Hodgkin's lymphoma
Homocysteine levels
Hypospadias
Idiopathic pulmonary fibrosis
IFN-related cytopeni
IgA levels
IgE levels
Inflammatory bowel disease
Insulin-like growth factors
Intracranial aneurysm
Iris color
Iron status markers
Ischemic stroke
Juvenile idiopathic arthritis
Keloid
Kidney stones
LDL cholesterol
Leprosy
Leptin receptor levels
Liver enzymes
Longevity
LP (a) levels
LpPLA(2) activity and mass
Lung cancer
Magnesium levels
Major mood disorders
Malaria
Male pattern baldness
Mammographic density
Matrix metalloproteinase levels
MCP-1
Melanoma
Menarche & menopause
Meningococcal disease
Metabolic syndrome
Migraine
Moyamoya disease
Multiple sclerosis
Myeloproliferative neoplasms
Myopia (pathological)
N-glycan levels
Narcolepsy
Nasopharyngeal cancer
Natriuretic peptide levels
Neuroblastoma
Nicotine dependence
Obesity
Open angle glaucoma
Open personality
Optic disc parameters
Osteoarthritis
Osteoporosis
Otosclerosis
Other metabolic traits
Ovarian cancer
Pancreatic cancer
Pain
Paget's disease
Panic disorder
Parkinson's disease
Periodontitis
Peripheral arterial disease
Personality dimensions
Phosphatidylcholine levels
Phosphorus levels
Photic sneeze
Phytosterol levels
Platelet count
Polycystic ovary syndrome
Primary biliary cirrhosis
Primary sclerosing cholangitis
PR interval
Progranulin levels
Progressive supranuclear palsy
Prostate cancer
Protein levels
PSA levels
Psoriasis
Psoriatic arthritis
Pulmonary funct. COPD
QRS interval
QT interval
Quantitative traits
Recombination rate
Red vs.non-red hair
Refractive error
Renal cell carcinoma
Renal function
Response to antidepressants
Response to antipsychotic therapy
Response to carbamazepine
Response to clopidogrel therapy
Response to hepatitis C treat
Response to interferon beta therapy
Response to metaformin
Response to statin therapy
Restless legs syndrome
Retinal vascular caliber
Rheumatoid arthritis
Ribavirin-induced anemia
Schizophrenia
Serum metabolites
Skin pigmentation
Smoking behavior
Speech perception
Sphingolipid levels
Statin-induced myopathy
Stroke
Sudden cardiac arrest
Suicide attempts
Systemic lupus erythematosus
Systemic sclerosis
T-tau levels
Tau AB1-42 levels
Telomere length
Testicular germ cell tumor
Thyroid cancer
Thyroid volume
Tooth development
Total cholesterol
Triglycerides
Tuberculosis
Type 1 diabetes
Type 2 diabetes
Ulcerative colitis
Urate
Urinary albumin excretion
Urinary metabolites
Uterine fibroids
Venous thromboembolism
Ventricular conduction
Vertical cup-disc ratio
Vitamin B12 levels
Vitamin D insuffiency
Vitiligo
Warfarin dose
Weight
White cell count
White matter hyperintensity
YKL-40 levels

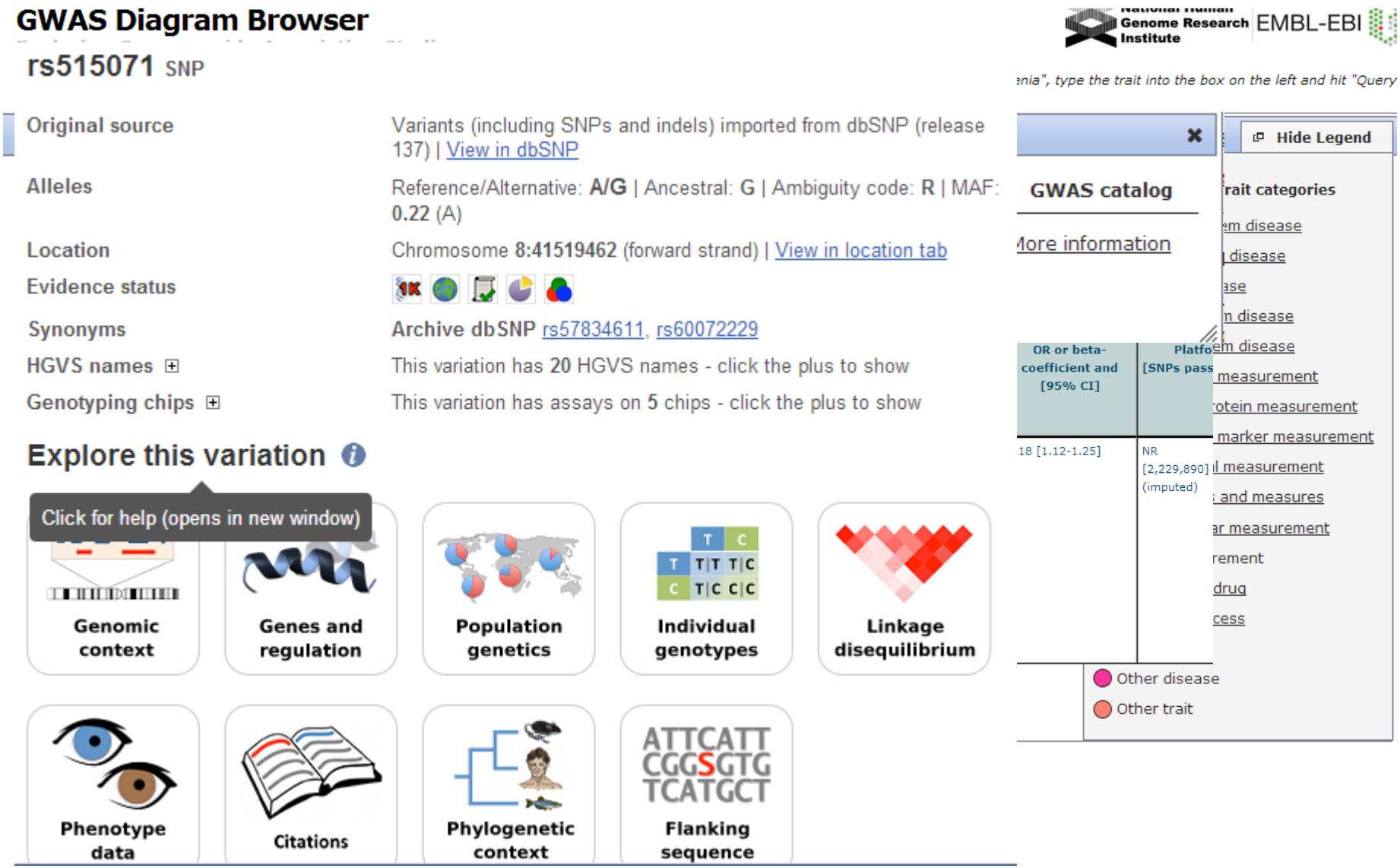
GWAS diagram automation

- Programmatic generation of the GWAS diagram from the GWAS/EFO knowledgebase
- Interactive diagram that can filtered by a number of criteria, e.g. to show only traits associated with a given disease
- Interactive traits (“dots”) that link directly into the catalogue
- New colour scheme with fewer colours representing higher-level trait categories, e.g. mental health disorders, cancers, cardio-vascular diseases

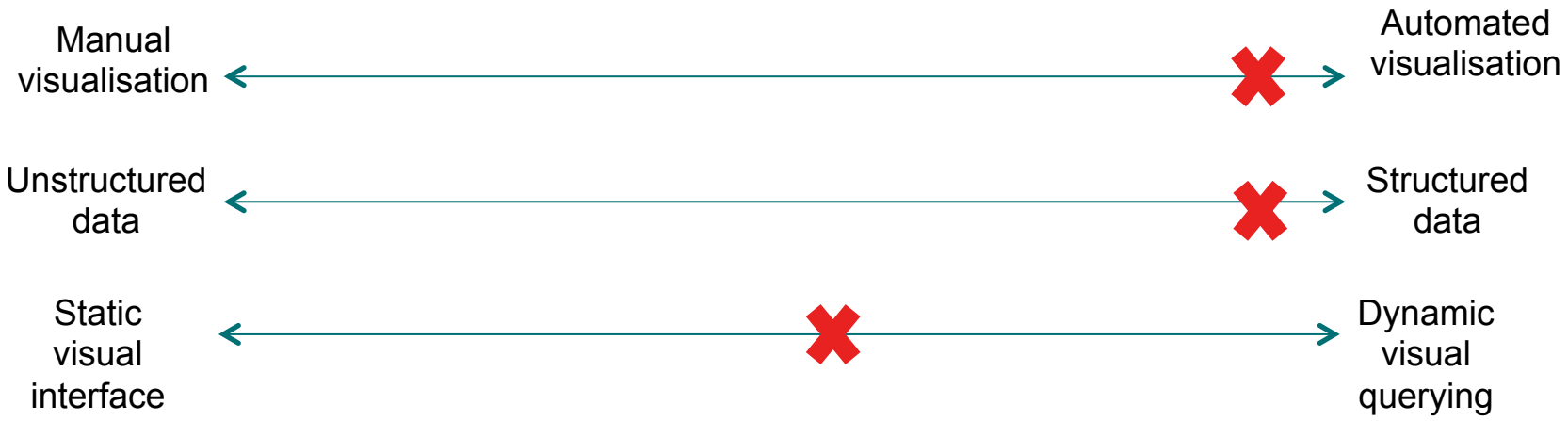
GWAS Visualisation www.ebi.ac.uk/fgpt/gwas



GWAS Data integration



Current status



- Web-application with back-end implemented in Java, running on an Apache Tomcat server
- Diagram generated in SVG
- Web-client – server communication via AJAX
- Client-side diagram manipulation in Javascript
- Hermit reasoner for classifying the OWL knowledgebase
- Continuous integration - monthly code releases, supporting data releases
- Code available on github, ontology available, all data available
- Component based Integration with NHGRI's Cold Fusion system for curation tracking

Summary

- Restructured GWAS catalogue data to allow querying beyond direct string matching
- Harmonised terms for all catalog content, re-mapped catalogue data for easier integration with other data sources
- Modelled the traits explicitly – e.g. disease and measurement
- Added new terms to the ontology to support the catalog
- Removed manual processing from catalogue visualisation
- Supported curators to choose terms during curation
- Used semantic web technologies for querying and visualisation of catalogue data

Future work

- Explore different resolution strategies for high-density regions
- Capture, model and query ethnicity information
- Better integration with genome browser
- Per study queries
- SNP level trait annotation and query
- Connect disease, phenotype and **assays**
  - ‘give me everything you have about diabetes’

Acknowledgements

- NHGRI
  - Peggy Hall
  - Lucia Hindorff
  - Heather Junkins
  - Kent Klemm
  - Darryl Leja
  - Teri Manolio

- EBI
  - Tony Burdett
  - Jon Ison
  - Simon Jupp
  - James Malone
  - Helen Parkinson
  - Joanella Morales
  - **Jackie MacArthur**
  - **Dani Welter**

NHGRI grant 3U41-HG006104-01S1
EMBL Core Funds