“Ecosystem for Collecting and Connecting Rare Disease Data”

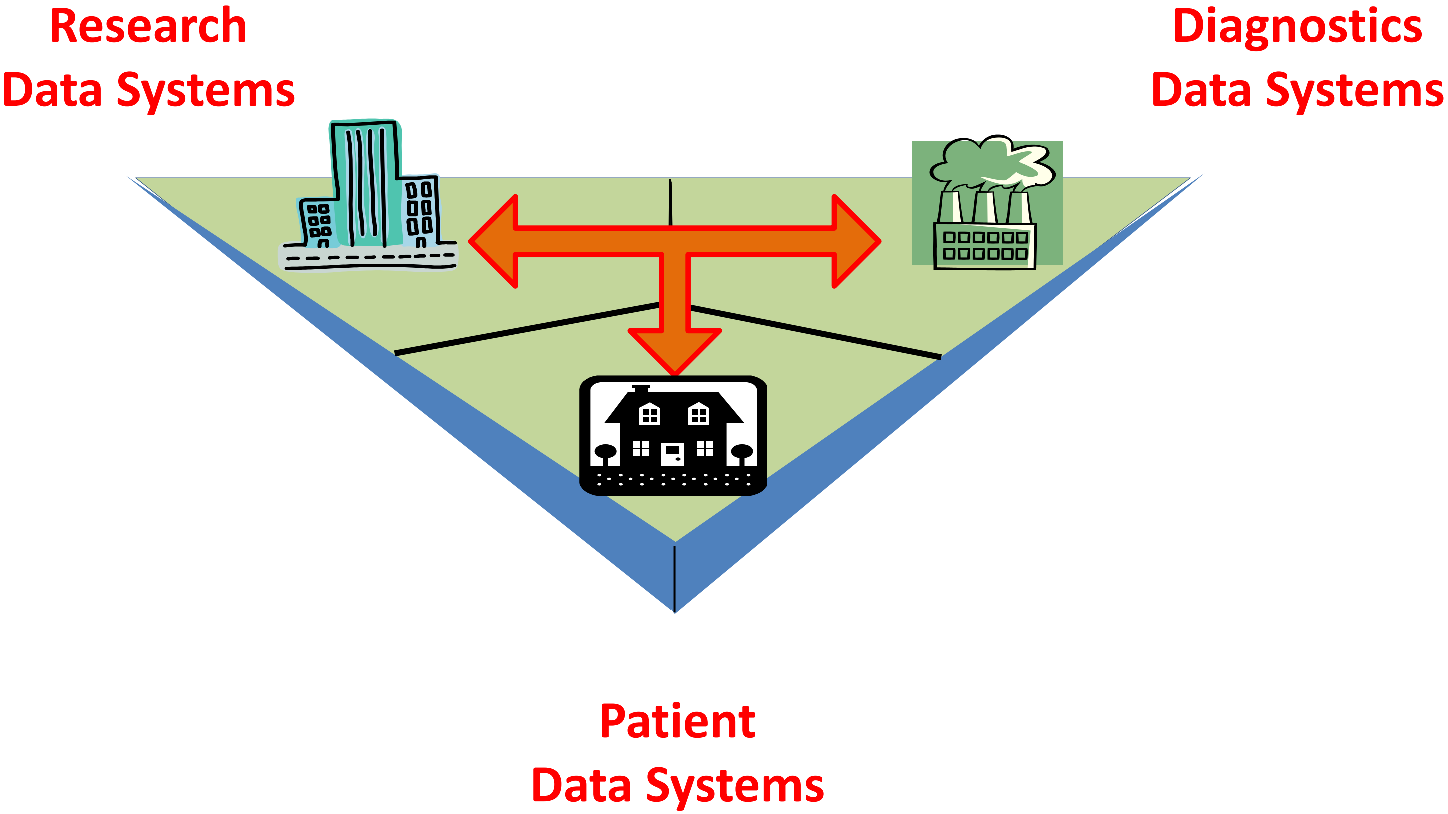
IRDiRC Conference

Dublin, 17 April 2013

Anthony J Brookes

University of Leicester

Ecosystem...

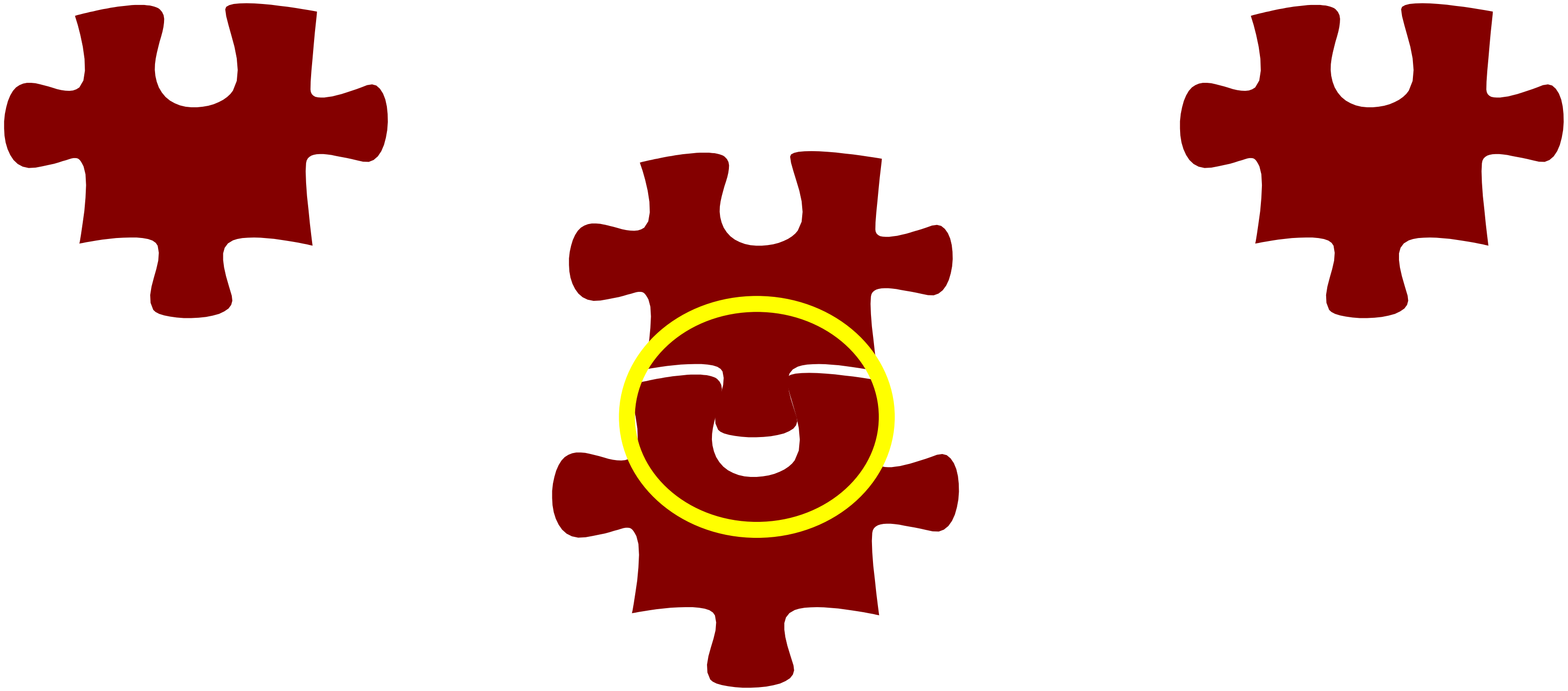


Research Data Systems

Diagnostics Data Systems

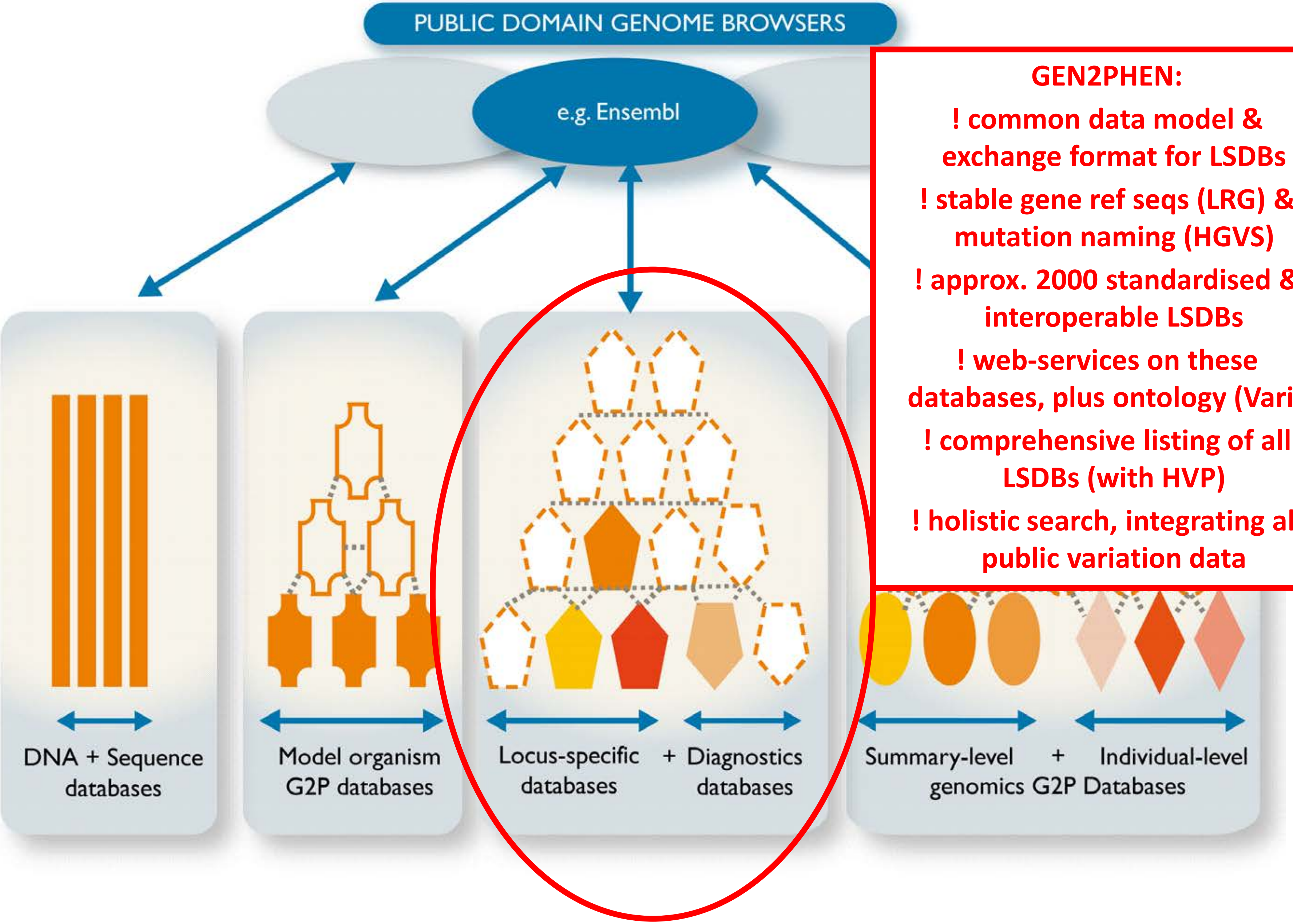
Patient Data Systems

Ecosystem...



So...

concentrate on collaboratively defining and promoting ‘connections’


PUBLIC DOMAIN GENOME BROWSERS
e.g. Ensembl
DNA + Sequence databases
Model organism G2P databases
Locus-specific databases
+ Diagnostics databases
Summary-level + Individual-level genomics G2P Databases

GEN2PHEN:

! common data model & exchange format for LSDBs

! stable gene ref seqs (LRG) & mutation naming (HGVS)

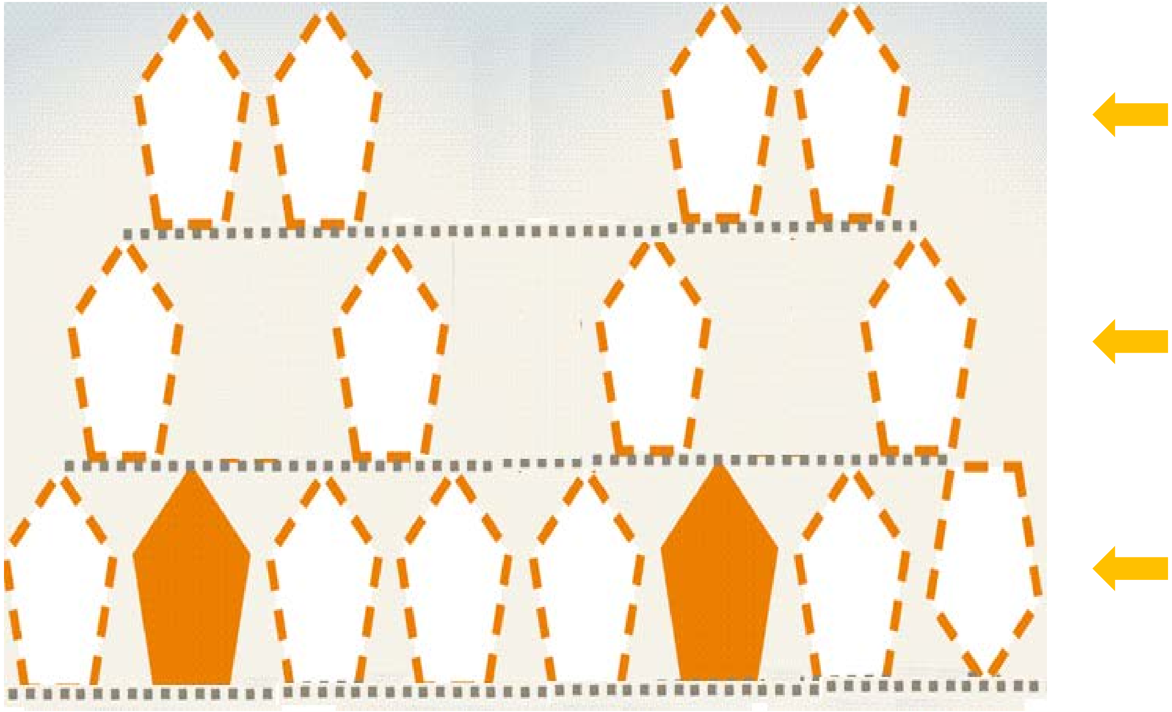
! approx. 2000 standardised & interoperable LSDBs

! web-services on these databases, plus ontology (Vario)

! comprehensive listing of all LSDBs (with HVP)

! holistic search, integrating all public variation data

Architecture



CENTRAL DBs

Safe core info, summaries

ClinVar, HGMD, JRC Registry

'INTEGRATION' DBs

Disease/ethnic focus, networks & consortia, external data federation

HVP nodes, DMuDB, PathoKB

SOURCE DBs

Expert curation, sensitive data

Research & diagnostic Labs, LSDBs, patient registries

ENTITY IDENTIFIERS

Data IDs

- The ‘Data Object Identifier’ (DOI) system, managed by DataCite. Covers a very broad concept of a ‘data object’ (much more than just traditional publications). Essential for creating the ‘web of data’.

Database IDs

- The BioDBCore project by which database IDs can be assigned. Essential if webservices are to start connecting resources effectively.

Human IDs

- The ‘Open Researcher Contributor Identifier’ (ORCID) system. Launched late 2013, has already issued many tens of thousands of ORCIDs. Soon to be a required author detail when submitting manuscripts. Removes ambiguity over all ‘contributors’, thereby enabling incentive/reward systems for data sharing, improved knowledge discovery options, and automation of data access control.

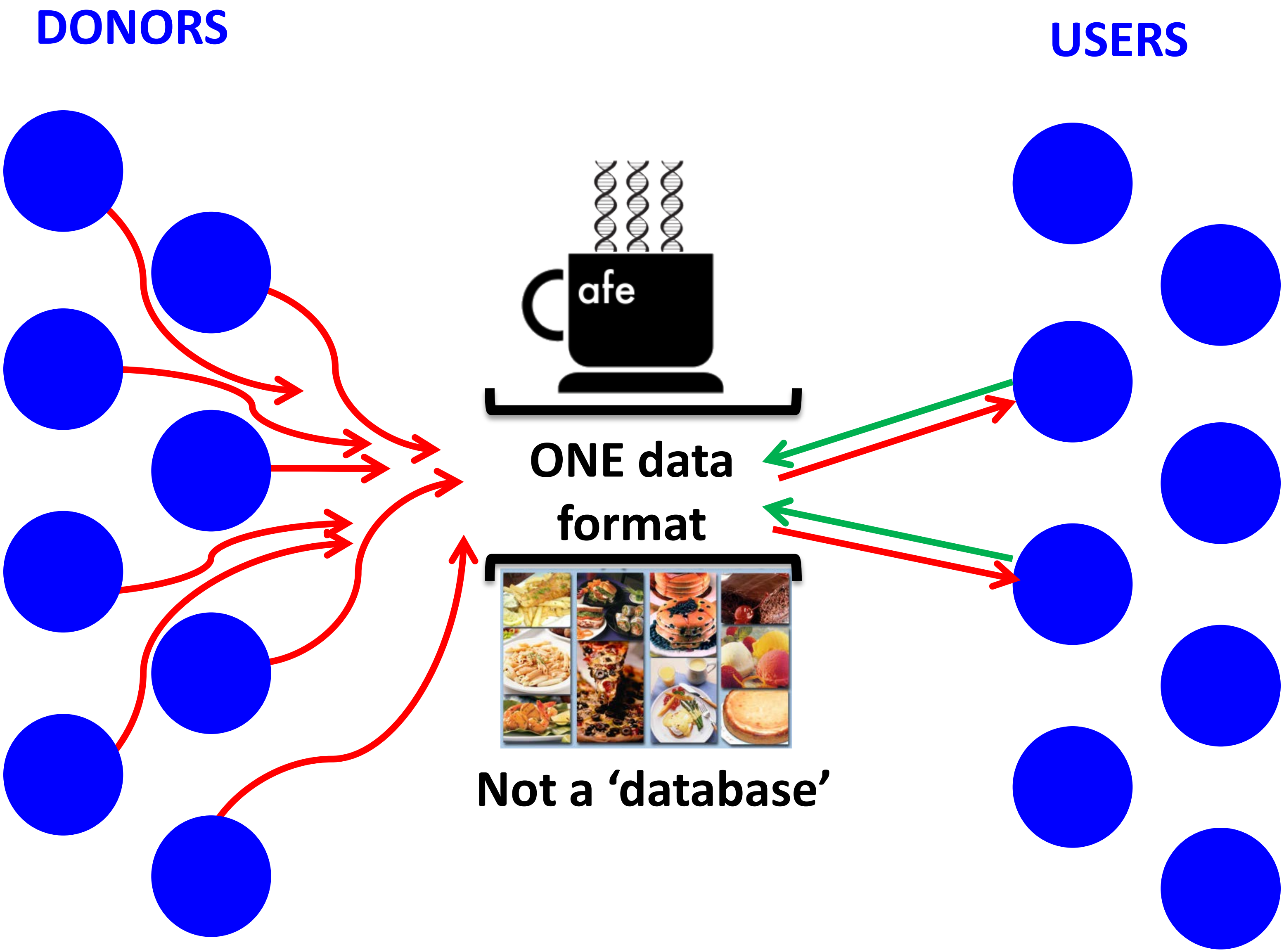
Biobank IDs

- Pilot system emerging from GEN2PHEN & BioShaRE, operated by P3G, as a basis for developing BioResource Impact Factor (BRIF) metrics.

Data Sharing & Access

Openly share the 'existence' rather than the 'substance' of the data
….thereafter variably manage data access

‘Cafe Variome’



DONORS

USERS

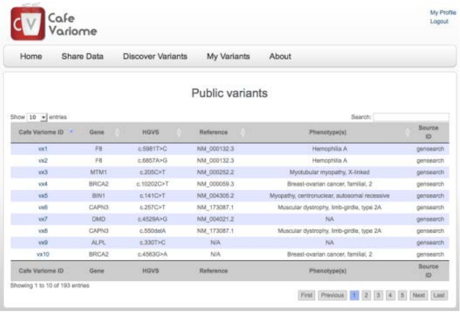
ONE data format

Not a ‘database’

Data Sharing Models (controlled access)

Open Access

Variants are made publically available for user



Public variants

Show 10 entries

Search:

Cafe Variome ID	Gene	HGVS	Reference	Phenotype(s)	Source ID
vx1	F8	c.5981T>C	NM_000132.3	Hemophilia A	gensearch
vx2	F8	c.6857A>G	NM_000132.3	Hemophilia A	gensearch
vx3	MTM1	c.205C>T	NM_000252.2	Myotubular myopathy, X-linked	gensearch
vx4	BRCA2	c.10202C>T	NM_000059.3	Breast-ovarian cancer, familial, 2	gensearch
vx5	BIN1	c.141C>T	NM_004305.2	Myopathy, centronuclear, autosomal recessive	gensearch
vx6	CAPN3	c.257C>T	NM_173087.1	Muscular dystrophy, limb-girdle, type 2A	gensearch
vx7	DMD	c.4529A>G	NM_004021.2	NA	gensearch
vx8	CAPN3	c.550delA	NM_173087.1	Muscular dystrophy, limb-girdle, type 2A	gensearch
vx9	ALPL	c.330T>C	N/A	NA	gensearch
vx10	BRCA2	c.4563G>A	N/A	Breast-ovarian cancer, familial, 2	gensearch
Cafe Variome ID	Gene	HGVS	Reference	Phenotype(s)	Source ID

Showing 1 to 10 of 193 entries

Export/view in multiple formats

Restricted Access

Enable permission to be conveniently sought from the data owner

?

Data owner easily approves/denies request. If approved, then data passed onto user

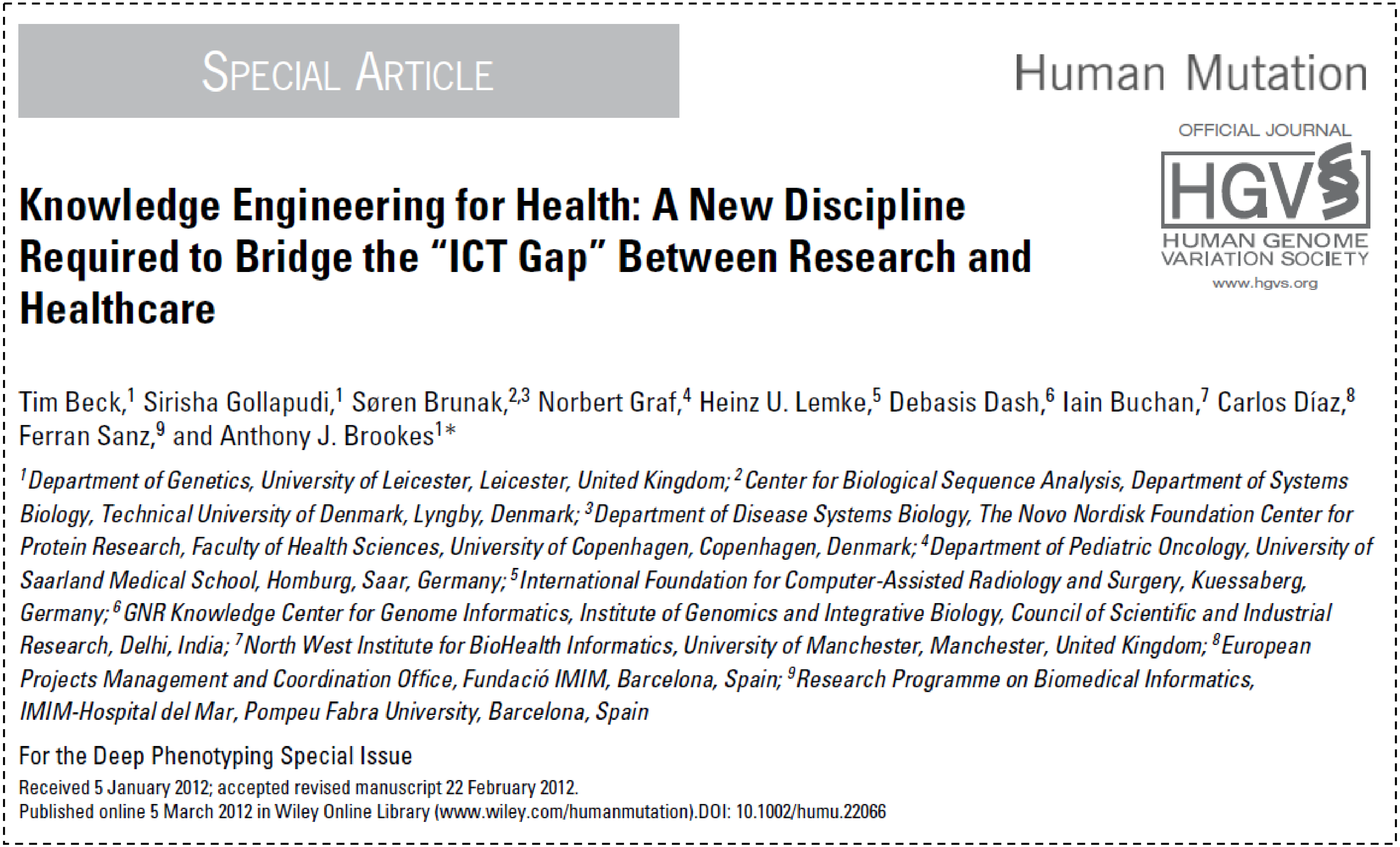
Linked Access

Variant is reported as a link to data source

Source DB/resource

Access managed via source db

Knowledge Engineering...



SPECIAL ARTICLE

Human Mutation

OFFICIAL JOURNAL

HGVS
HUMAN GENOME
VARIATION SOCIETY
www.hgvs.org

Knowledge Engineering for Health: A New Discipline Required to Bridge the “ICT Gap” Between Research and Healthcare

Tim Beck,[1] Sirisha Gollapudi,[1] Søren Brunak,[2,3] Norbert Graf,[4] Heinz U. Lemke,[5] Debasis Dash,[6] Iain Buchan,[7] Carlos Díaz,[8] Ferran Sanz,[9] and Anthony J. Brookes[1]*

[1] Department of Genetics, University of Leicester, Leicester, United Kingdom; [2] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark; [3] Department of Disease Systems Biology, The Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark; [4] Department of Pediatric Oncology, University of Saarland Medical School, Homburg, Saar, Germany; [5] International Foundation for Computer-Assisted Radiology and Surgery, Kuessaberg, Germany; [6] GNR Knowledge Center for Genome Informatics, Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research, Delhi, India; [7] North West Institute for BioHealth Informatics, University of Manchester, Manchester, United Kingdom; [8] European Projects Management and Coordination Office, Fundació IMIM, Barcelona, Spain; [9] Research Programme on Biomedical Informatics, IMIM-Hospital del Mar, Pompeu Fabra University, Barcelona, Spain

For the Deep Phenotyping Special Issue

Received 5 January 2012; accepted revised manuscript 22 February 2012.
Published online 5 March 2012 in Wiley Online Library (www.wiley.com/humanmutation).DOI: 10.1002/humu.22066

- presentation & discussion at many international meetings and forums
    - 1/2 day workshop as satellite to ESHG (6 invited speakers)
    - workshop session at MIE2011 (3 invited speakers, audience discussion)
    - I-Health 2011 workshop in Brussels, 3-4 Oct 2011

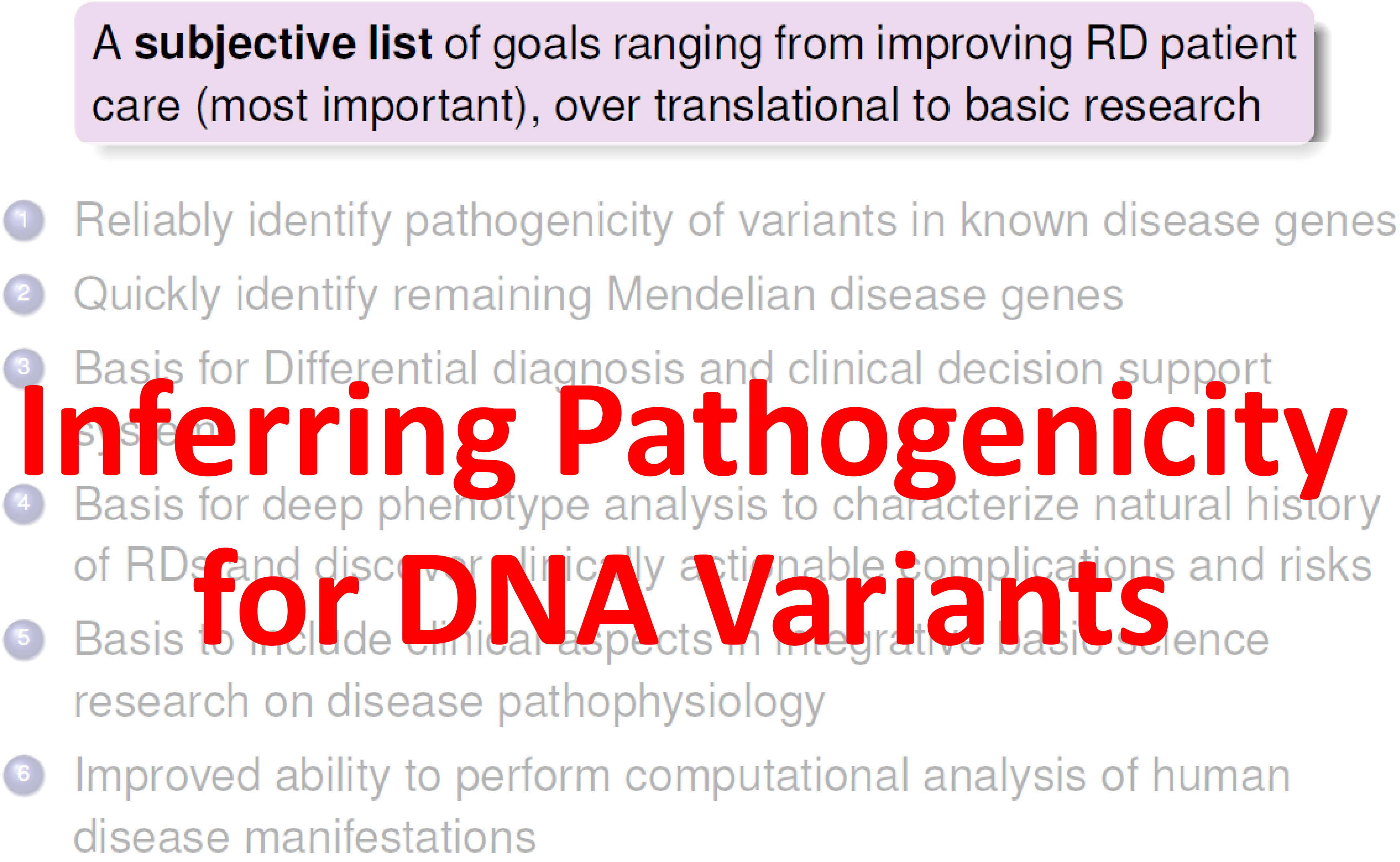
- growing community, currently >150 academics, companies, healthcare providers

I4Health network

Integration and Interpretation of Information for Individualised Healthcare

http://www.i4health.eu/

Inferring Pathogenicity for DNA Variants



A **subjective list** of goals ranging from improving RD patient care (most important), over translational to basic research

1. Reliably identify pathogenicity of variants in known disease genes
2. Quickly identify remaining Mendelian disease genes
3. Basis for Differential diagnosis and clinical decision support system
4. Basis for deep phenotype analysis to characterize natural history of RDs and discover clinically actionable complications and risks
5. Basis to include clinical aspects in integrative basic science research on disease pathophysiology
6. Improved ability to perform computational analysis of human disease manifestations

HOW TO INFER PATHOGENICITY…

- Allele frequency in controls (matched population?)
- Relevant publication (listed in HGMD)
- Presence or absence in variant databases (LSDB, dbSNP, ClinVar)
- Co-segregation with the disease in the family
- Cross-Species conservation
- Protein structure predictions
- *In silico* prediction of pathogenic effect
  (e.g., Align GVGD, PolyPhen-2, SIFT, MutationTaster)
- *In silico* splice site prediction
  (e.g., SSF, MaxEnt, NNSPLICE, GeneSplicer, HSF)
- Functional Studies - human context
- Functional Studies - model organism context

PATHOGENICITY

- ‘Pathogenicity’ = two related concepts:
(a) whether a variant has ‘caused’ a phenotype in a particular patient/family
(b) whether a variant can ‘cause’ a phenotype in anyone in a population

- ‘Pathogenicity Score, or Non-Irrelevance Score’
= <u>degree of certainty</u> that a genetic variant is <u>not completely benign</u>
irrespective of *e.g., environment, nutrition, gender, age, genetic/metabolome/epigenetic background, zygosity, copy number, mosaicism, etc*

..also

- ‘Penetrance Score’ = <u>range and distribution of likelihood</u> that phenotype will result, in a specified situation (e.g., age, gender, population, environment…)

- ‘Expressivity Score’ = <u>range and distribution of severity</u> of phenotype caused, in a specified situation (e.g., age, gender, population, environment…)

- Evidence base = types, reliability, and quantitative weighting of items of evidence that inform the pathogenicity metrics

- Phenotype = pathogenicity is only meaningful in the context of a properly define phenotype

- Actionability = determined by all extremes of ‘pathogenicity’, ‘penetrance’ & ‘expressivity’

GEN2PHEN Partners (www.gen2phen.org)

Academic

A.J.Brookes, R.Dalgleish	University of Leicester	UK
P.Flicek, H.Parkinson	European Molecular Biology Laboratory	Germany
C.Díaz	Fundació IMIM	Spain
J.denDunnen	Leiden University Medical Centre	Netherlands
C.Béroud	Inst Natl de la Santé et de la Recherche Méd	France
A.Cambon-Thomsen	Inst Natl de la Santé et de la Recherche Méd	France
J-E.Litton	Karolinska Institute	Sweden
G.Potamias	Foundation for Research & Technology	Greece
G.Patrinos	University of Patras	Greece
M.Lathrop	Centre National de Génotypage	France
J.Muilu	University of Helsinki	Finland
J.L.Oliveira	University of Aveiro – IEETA	Portugal
D.Dash	Institute of Genomics and Integrative Biology	India
L.Yip	Swiss Institute of Bioinformatics	Switzerland
A.Devereau	University of Manchester	UK
M.Swertz	Groningen University Medical Centre	Netherlands
M.Vihinen	University of Tampere	Finland

SMEs

A.Kel	BioBase GmbH	Germany
H.Gudbjartsson	deCODE genetics	Iceland
D.Atlan	PhenoSystems	Belgium
T.Kanninen	Biocomputing Platforms	Finland

Previous

H.Lehvaslaiho	University of Western Cape	South Africa

Acknowledgments

- GEN2PHEN Partners
- Visionaries:
  David Atlan, Segolène Aymé, Anne Cambon-Thomsen, Andrew Devereau, Carlos Díaz, Johan den Dunnen, Xavier Estivill , Matthew Hurles, Marie-Christine Jaulent, Gert Matthijs, Barend Mons, Georges de Moor, Yves Moreau, Juha Muilu, Peter Robinson, Patrick Ruch, Paul Schofield, Morris Swertz, David Voets, Steven van Vooren
- My team:
  Robert Free, Rob Hastings, Adam Webb, Tim Beck, Sirisha Gollapudi, Gudmundur Thorisson, Owen Lancaster

"Data-to-Knowledge-for-Practice" (DKP) Center

HGVbaseG2P has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 200754 - the GEN2PHEN project.