

Centro Nacional de Análisis Genómico

Where are the Bottlenecks of Genome Analysis Today?

Teratec

Ecole Polytechnique, Palaiseau, F

Ivo Glynne Gut

29.06.2016

cnag

centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

 **CRG**
Centre
for Genomic
Regulation

The genomehenge

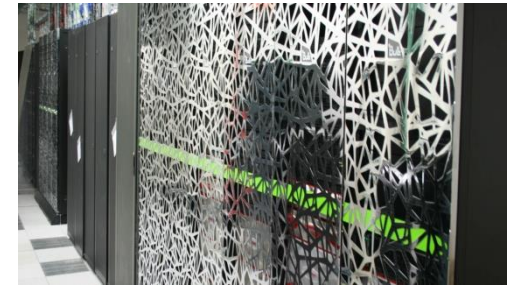


Sequencing capacity

- >1000 Gbases/day = 9-10 human genomes per day at 30x coverage

Equipment

- 12 Illumina HiSeq2000/2500/4000
- 1 Illumina MiSeq
- 4 Illumina cBots
- 3 Oxford Nanopore Minions
- Caliper/Eppendorf liquid handling robotics
- Fluidigm C1
- Bull 3500 core cluster super computer
- Maxeler Data Flow Engine
- 200 Tflops
- 7.5 petabyte disc space/tape
- Barcelona Supercomputing Center (10 x 10 Gb/s)



Copyright 2005, Barcelona Supercomputing Center - BSC

Size of a Human Genome Sequence

- Twice 3.200.000.000 nucleotides
- One human genome requires sequencing 100.000.000.000 nucleotides (T,C,G or A)
- If one nucleotide is a corn of rice this would fill TWO Olympic size swimming pools

Sequencing of a Human Genome

- 1.000.000.000 fragments of 100 nucleotides
- Alignment against the reference sequence of the human genome
- Identification of the deviations from the reference
- Computational requirements ~ 2000 CPUh per human genome

PERSPECTIVES

International network of cancer genome projects

The International Cancer Genome Consortium*

The International Cancer Genome Consortium (ICGC) was launched to coordinate large-scale cancer genome studies in tumours from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe. Systematic studies of more than 25,000 cancer genomes at the genomic, epigenomic and transcriptomic levels will reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies.



Nature 464, 993-998 (15.04.2010)

cnag

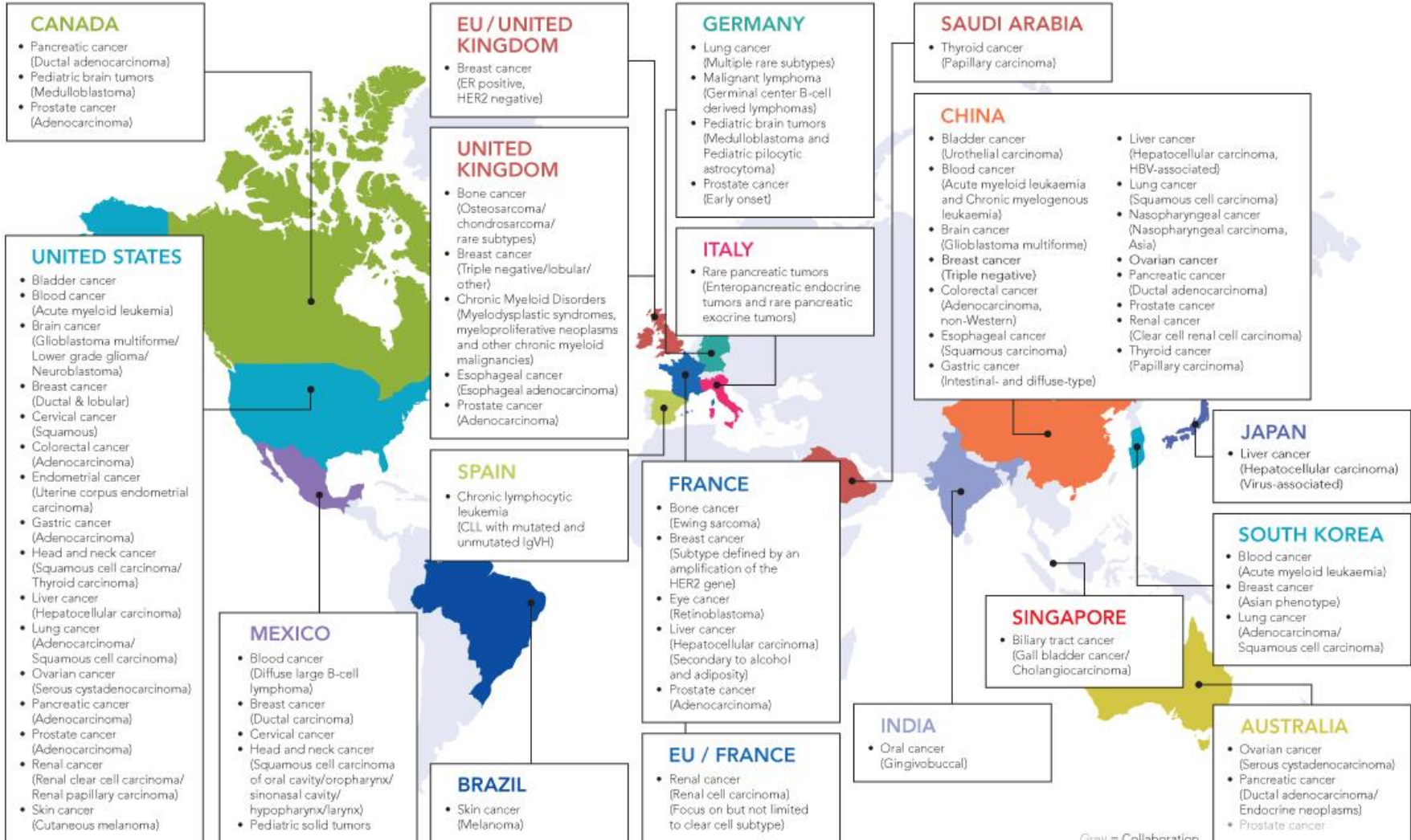
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico



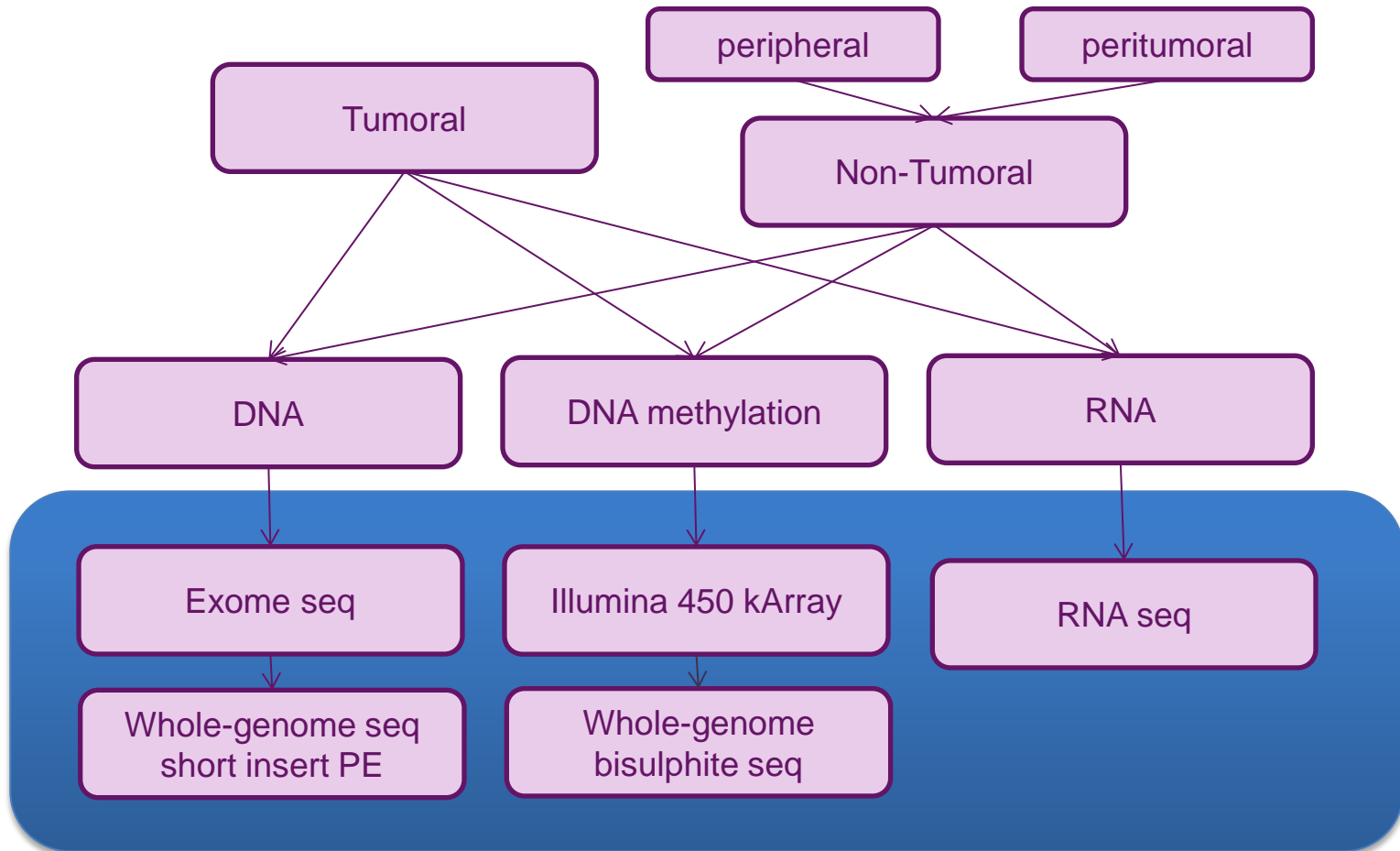


International Cancer Genome Consortium

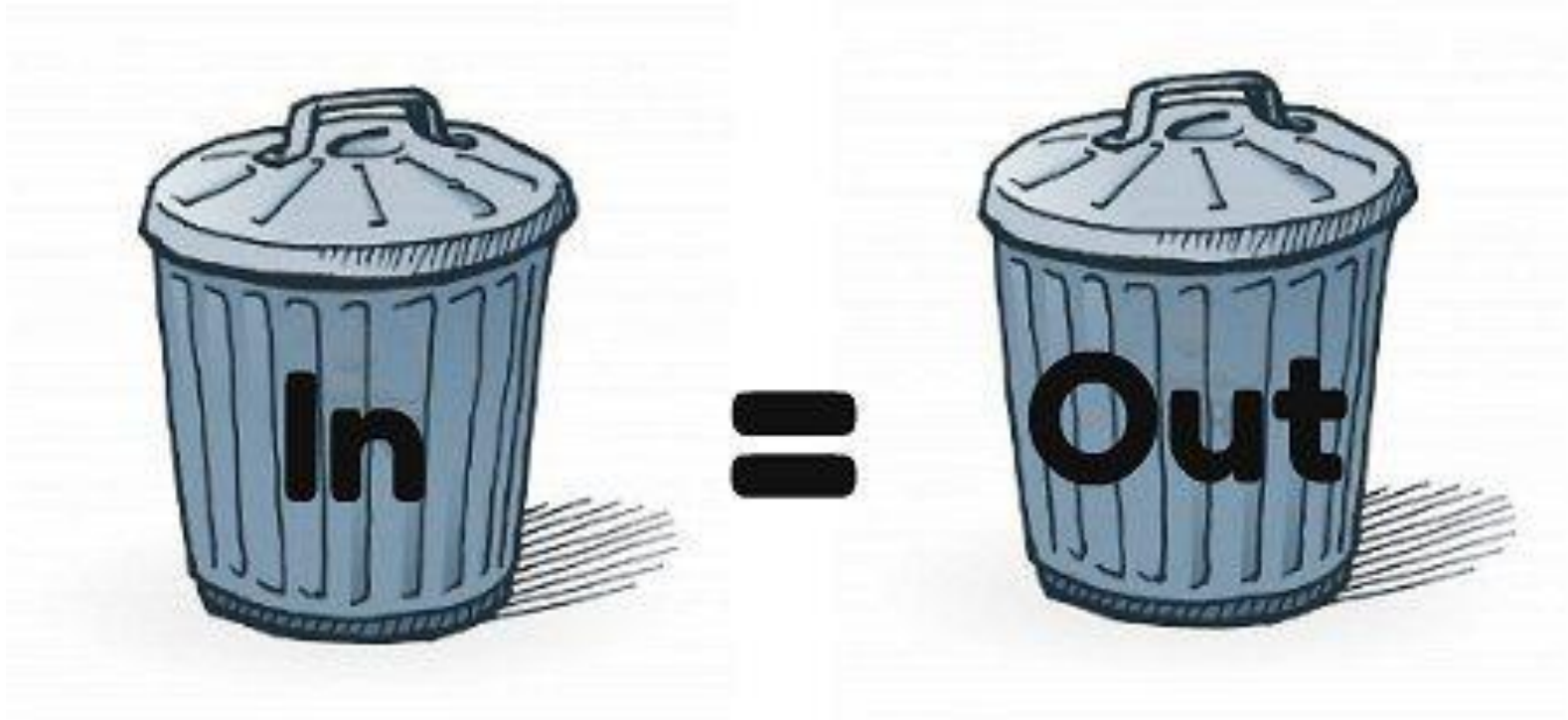
83 Projects/18 Countries



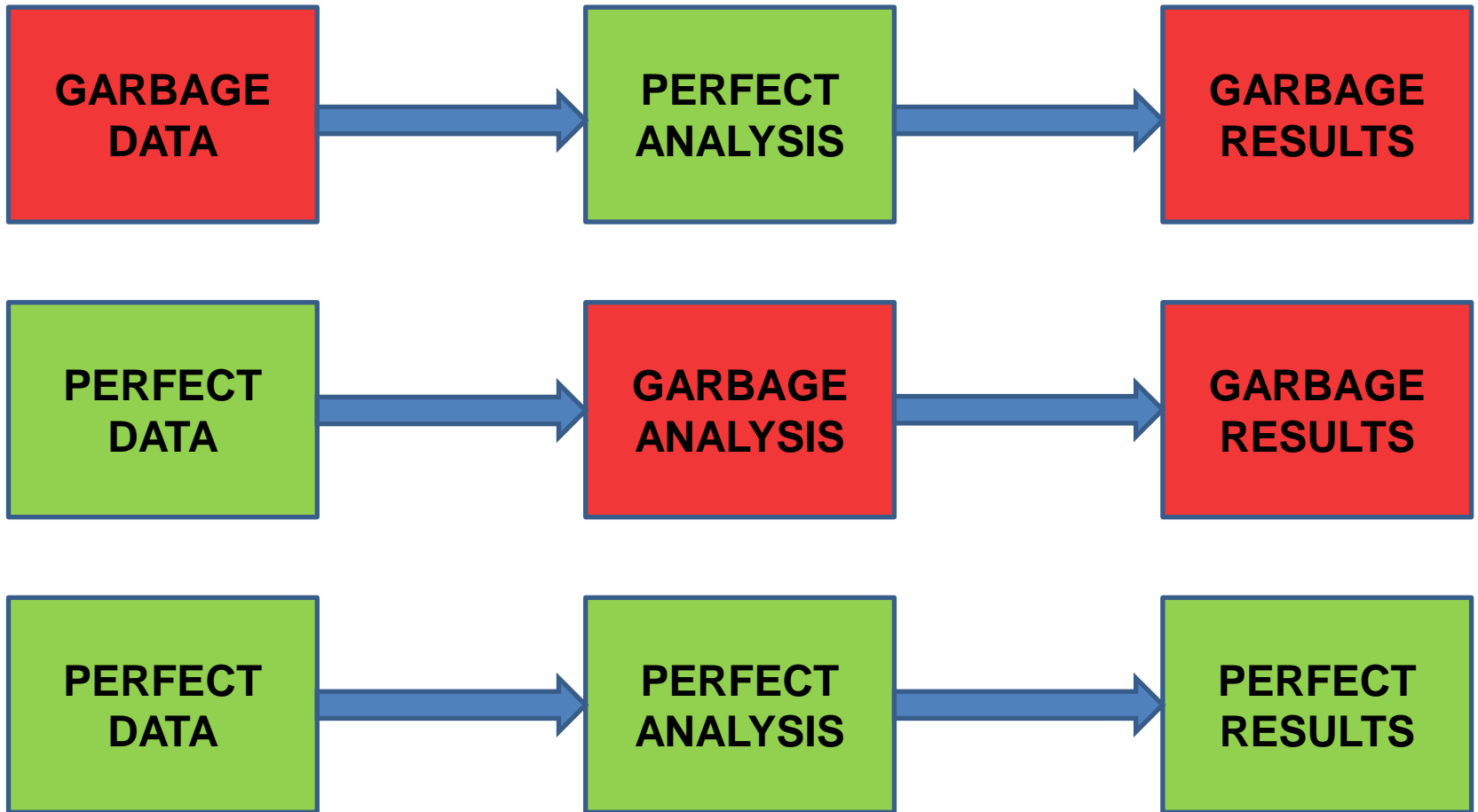
Sampling structure



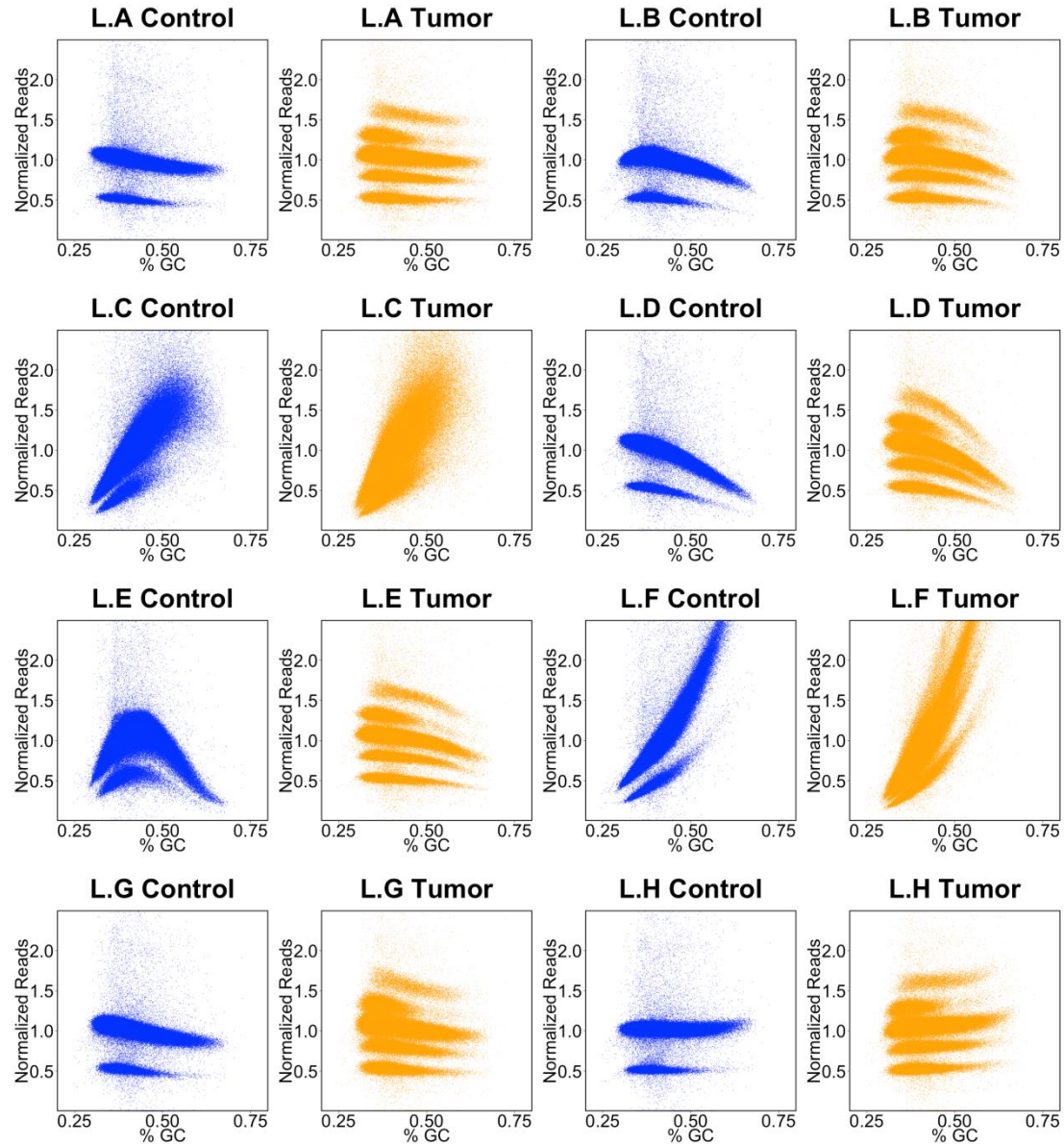




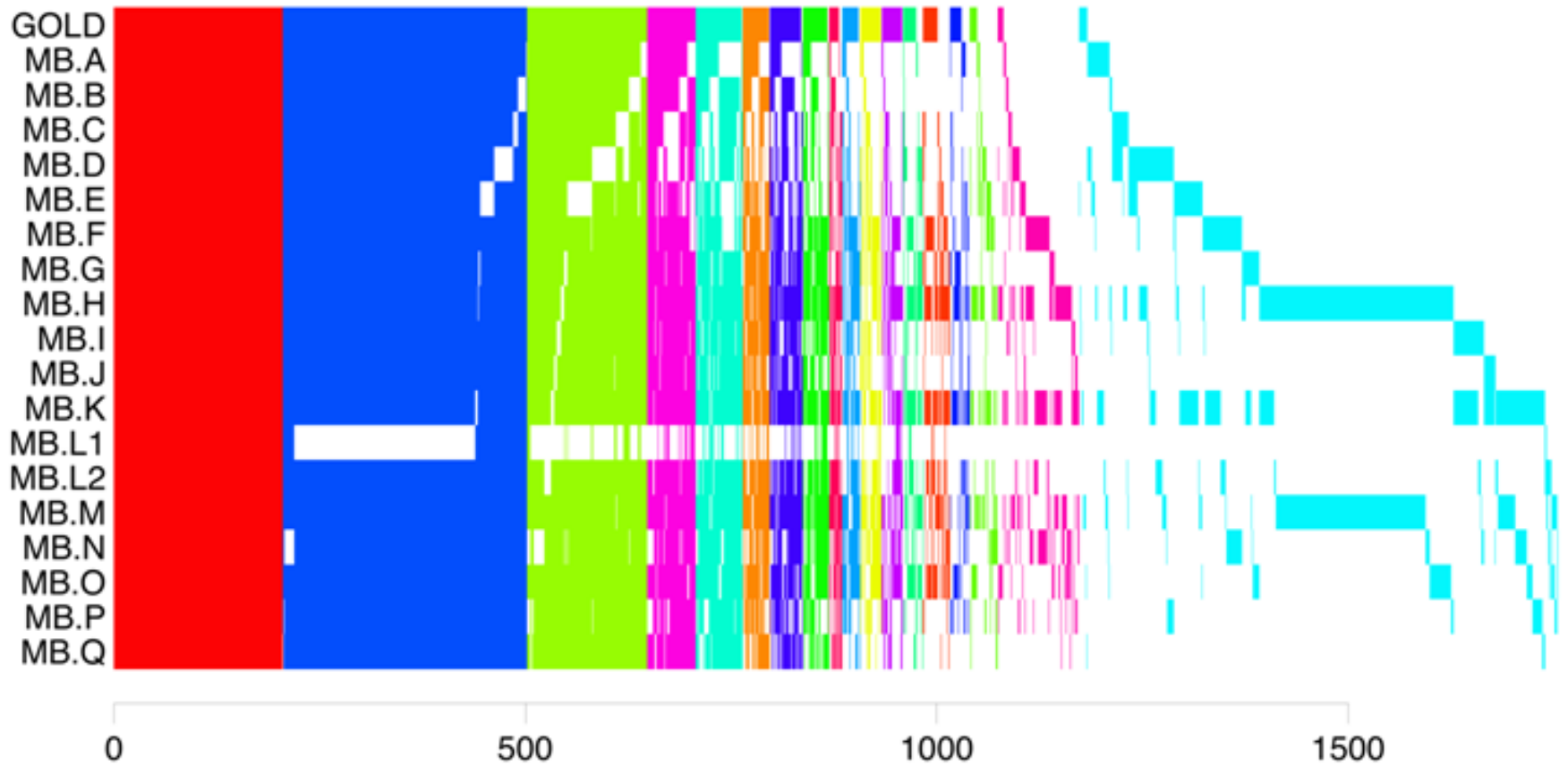
Garbage In – Garbage Out Paradigm



GC bias of different libraries



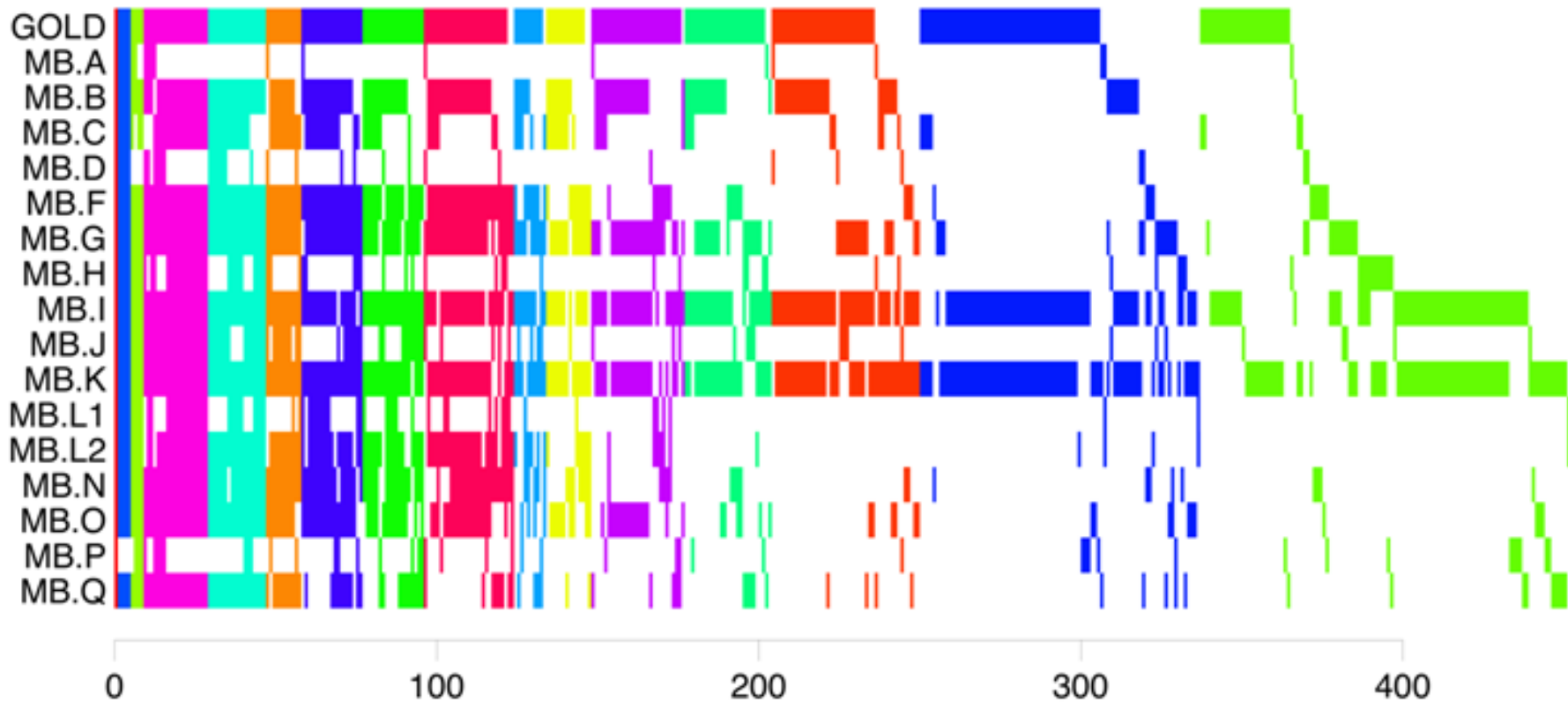
BM1.2 Somatic Single-base Mutations (SSM)



18 submissions



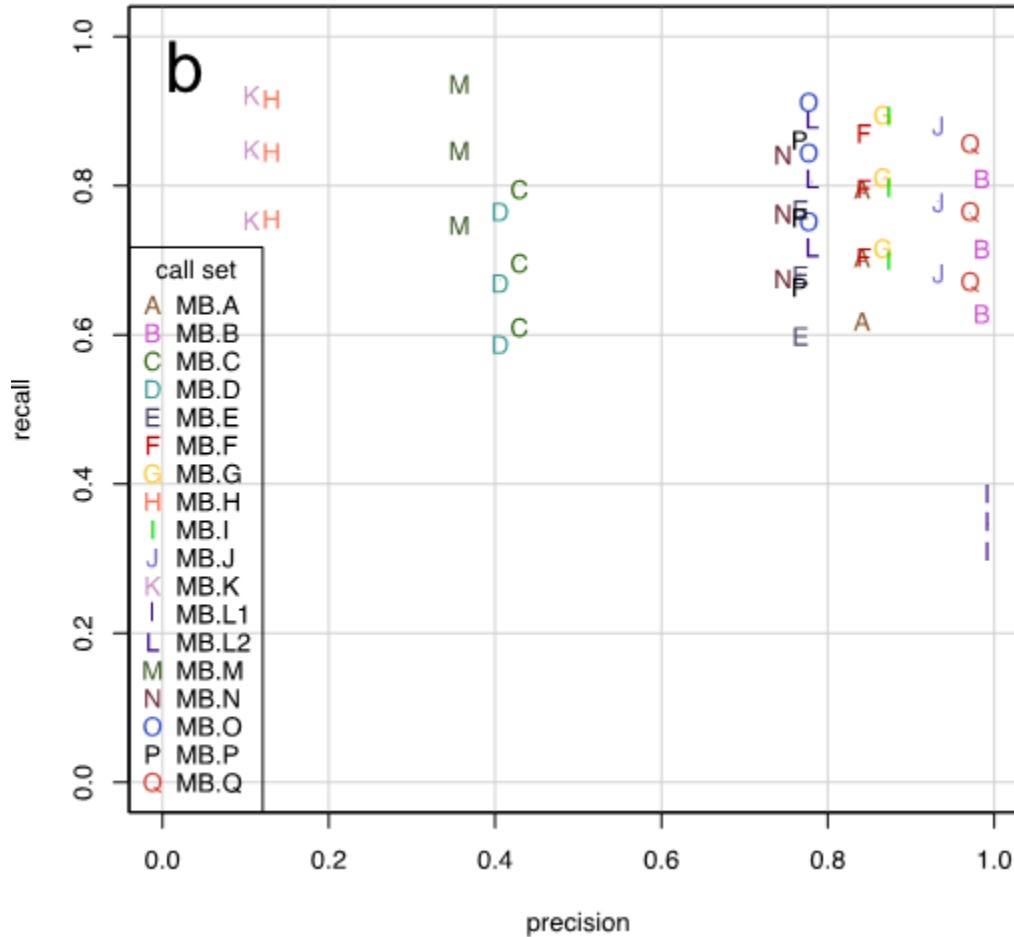
Somatic Insertion/deletion Mutations (SIM)



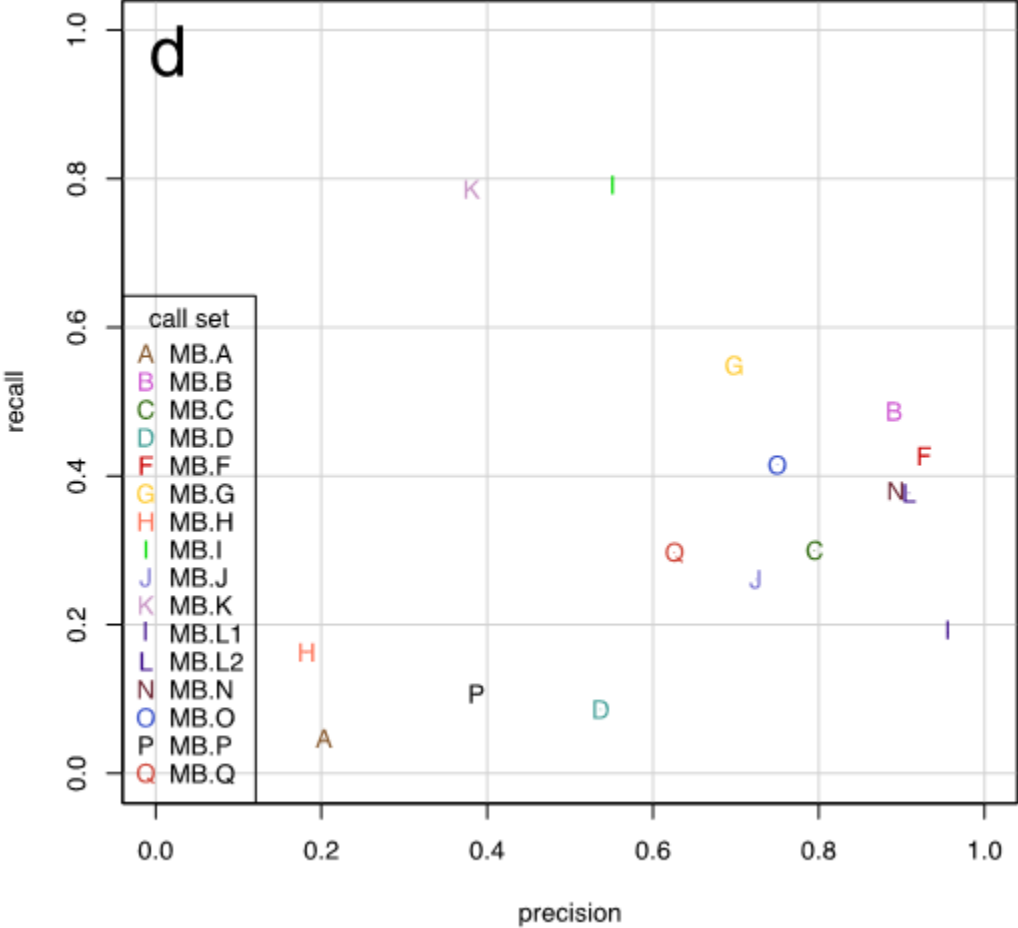
16 submissions



BM 1.2 SSMs Sensitivity and Specificity



SIMs Sensitivity and Specificity





ARTICLE

Received 16 Jun 2015 | Accepted 23 Oct 2015 | Published xx xxx 2015

DOI: 10.1038/ncomms10001

OPEN

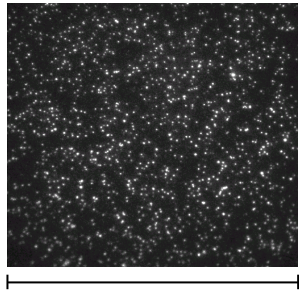
A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing

Tyler S. Alioto^{1,2,*}, Ivo Buchhalter^{3,4,*}, Sophia Derdak^{1,2}, Barbara Hutter⁴, Matthew D. Eldridge⁵, Eivind Hovig^{6,7}, Lawrence E. Heisler⁸, Timothy A. Beck⁸, Jared T. Simpson⁸, Laurie Tonon⁹, Anne-Sophie Sertier⁹, Ann-Marie Patch^{10,11}, Natalie Jäger^{3,12}, Philip Ginsbach³, Ruben Drews³, Nagarajan Paramasivam³, Rolf Kabbe³, Sasithorn Chotewutmontri¹³, Nicole Diessl¹³, Christopher Previti¹³, Sabine Schmidt¹³, Benedikt Brors⁴, Lars Feuerbach⁴, Michael Heindorf⁴, Susanne Gröbner¹⁴, Andrey Korshunov¹⁵, Patrick S. Tarpey¹⁶, Adam P. Butler¹⁶, Jonathan Hinton¹⁶, David Jones¹⁶, Andrew Menzies¹⁶, Keiran Raine¹⁶, Rebecca Shepherd¹⁶, Lucy Stebbings¹⁶, Jon W. Teague¹⁶, Paolo Ribeca^{1,2}, Francesc Castro Giner^{1,2}, Sergi Beltran^{1,2}, Emanuele Raineri^{1,2}, Marc Dabad^{1,2}, Simon C. Heath^{1,2}, Marta Gut^{1,2}, Robert E. Denroche⁸, Nicholas J. Harding⁸, Takafumi N. Yamaguchi⁸, Akihiro Fujimoto¹⁷, Hidewaki Nakagawa¹⁷, Victor Quesada¹⁸, Rafael Valdés-Mas¹⁸, Sigve Nakken⁶, Daniel Vodák^{6,19}, Lawrence Bower⁵, Andrew G. Lynch⁵, Charlotte L. Anderson^{5,20}, Nicola Waddell^{10,11}, John V. Pearson^{10,11}, Sean M. Grimmond^{10,21}, Myron Peto²², Paul Spellman²², Minghui He²³, Cyriac Kandoth²⁴, Semin Lee²⁵, John Zhang^{25,26}, Louis Létourneau²⁷, Singer Ma²⁸, Sahil Seth²⁶, David Torrents²⁹, Liu Xi³⁰, David A. Wheeler³⁰, Carlos López-Otin¹⁸, Elias Campo³¹, Peter J. Campbell¹⁶, Paul C. Boutros^{9,32}, Xose S. Puente¹⁸, Daniela S. Gerhard³³, Stefan M. Pfister^{14,34}, John D. McPherson^{8,32}, Thomas J. Hudson^{8,32,35}, Matthias Schlesner³, Peter Lichter^{36,37}, Roland Eils^{3,37,38,39,*}, David T.W. Jones^{40,*} & Ivo G. Gut^{1,2,*}

As whole-genome sequencing for cancer genome analysis becomes a clinical tool, a full understanding of the variables affecting sequencing analysis output is required. Here using tumour-normal sample pairs from two different types of cancer, chronic lymphocytic leukaemia and medulloblastoma, we conduct a benchmarking exercise within the context of the International Cancer Genome Consortium. We compare sequencing methods, analysis pipelines and validation methods. We show that using PCR-free methods and increasing sequencing depth to $\sim 100\times$ shows benefits, as long as the tumour:control coverage ratio remains balanced. We observe widely varying mutation call rates and low concordance among analysis pipelines, reflecting the artefact-prone nature of the raw data and lack of standards for dealing with the artefacts. However, we show that, using the benchmark mutation set we have created, many issues are in fact easy to remedy and have an immediate positive impact on mutation detection accuracy.

Value of data and its footprint

Image



100um
Random array
of clusters

fastq

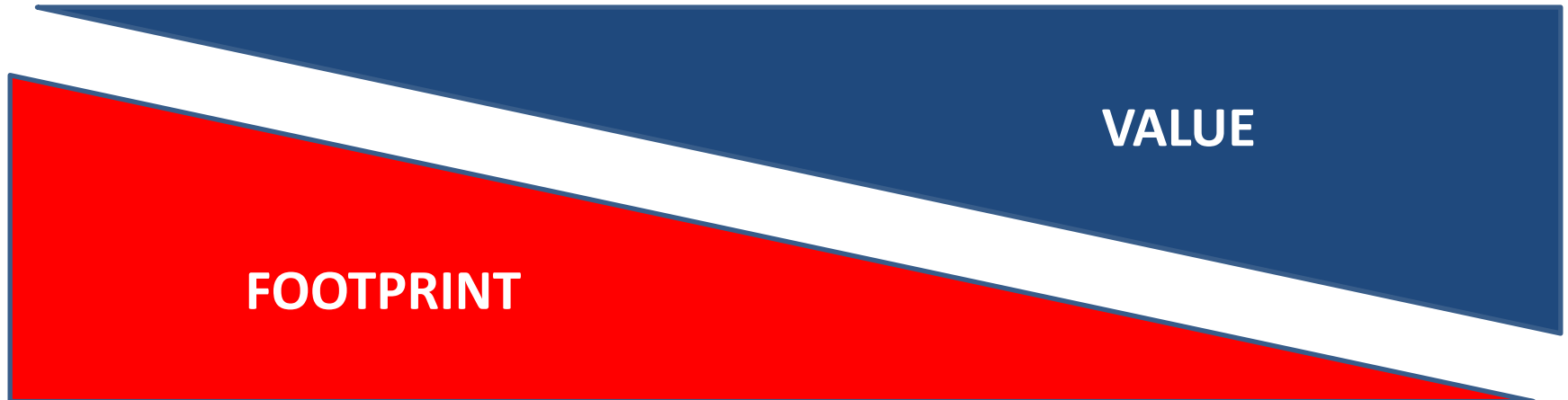
```
atcgagctttagagaggat
ggcgatttagtagcagggg
atcgagctttccttttaggat
ctcaagctttaggtaaggcc
ctcaagggaaagtaagttc
```

bam

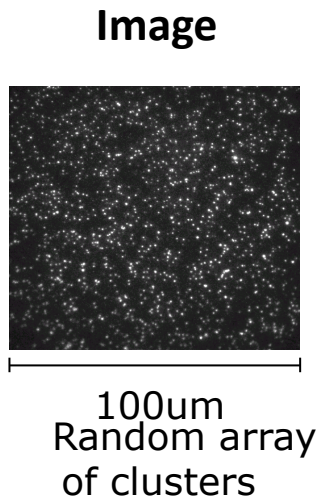
```
atcgagctttagagaggat
tcgagctttagagaggata
cgagctttagagaggataa
gagctttagagaggataaa
agctttagagaggataaag
```

vcf

```
t/c
t/-
a/g
t/c
t/a
```



Value of data and its footprint




fastq

```
atcgagctttagagaggat
ggcgatttagtagcagggg
atcgagctttccttttaggat
ctcaagctttaggtaaggcc
ctcaagggaagtaagttc
```

bam

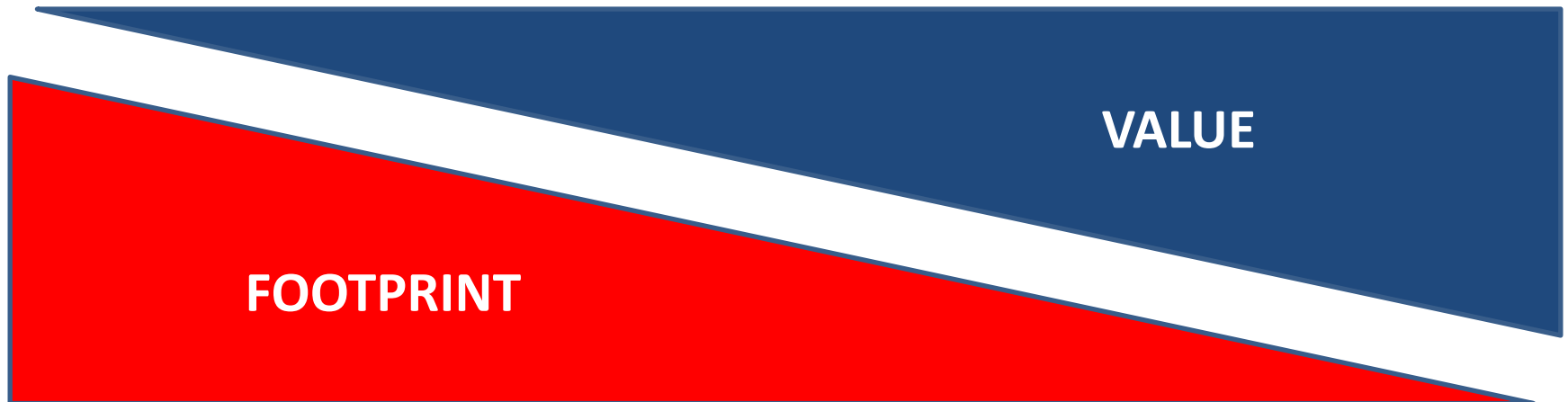
```
atcgagctttagagaggat
tcgagctttagagaggata
cgagctttagagaggataa
gagctttagagaggataaa
agctttagagaggataaag
```

Phenotype



vcf

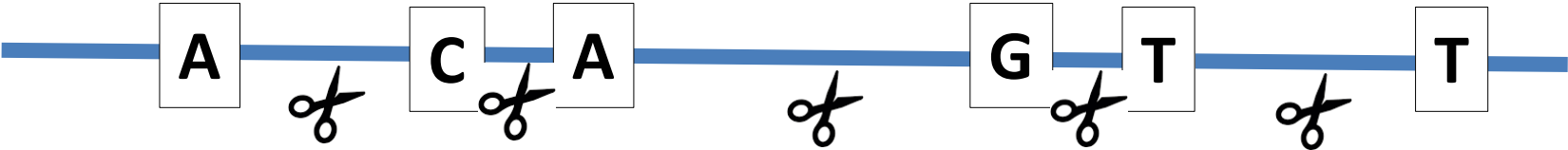
```
t/c
t/-
a/g
t/c
t/a
```



Identifiability



Identifiability



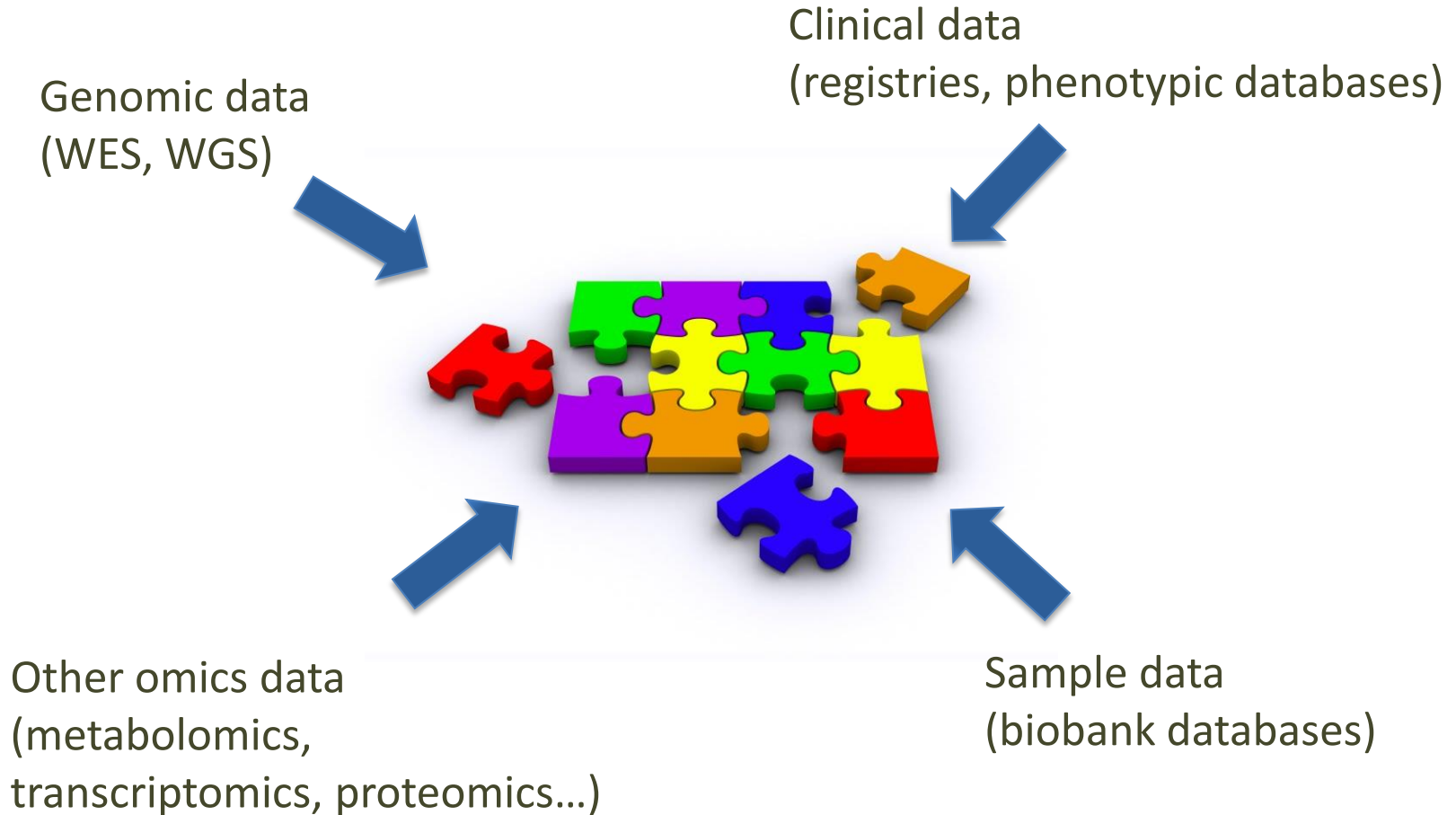
RD-CONNECT UPDATE:
MARCH 2016

Rare Diseases

- 1 in 2000 people of the general population
- ~ 9000 rare diseases
- ~ 12% of the population is affected by a rare disease

- Rare disease patients have an interest that their data is used/shared

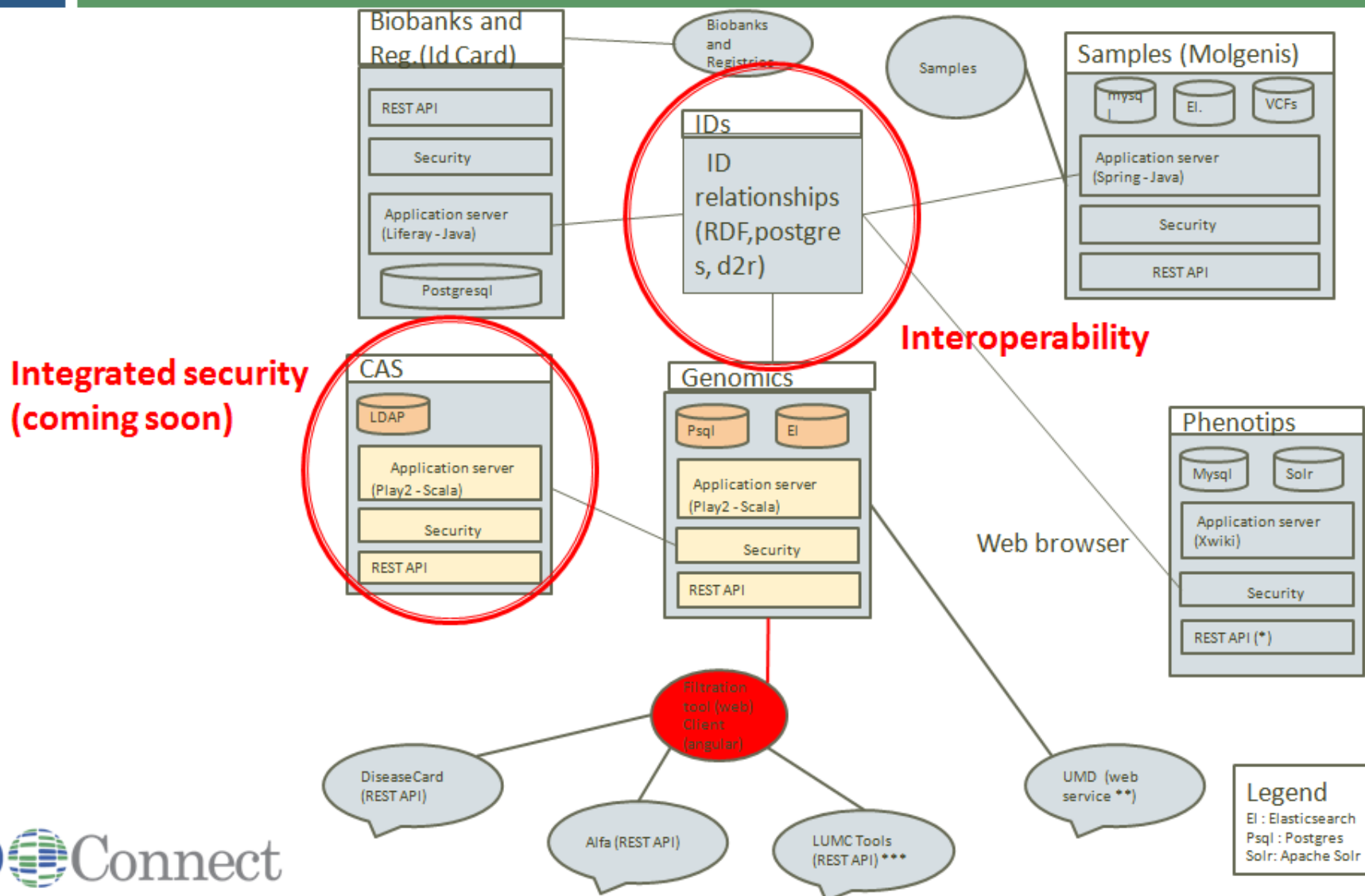
Data in the RD-Connect platform





Whole RD-Connect Platform Architecture Overview 2016

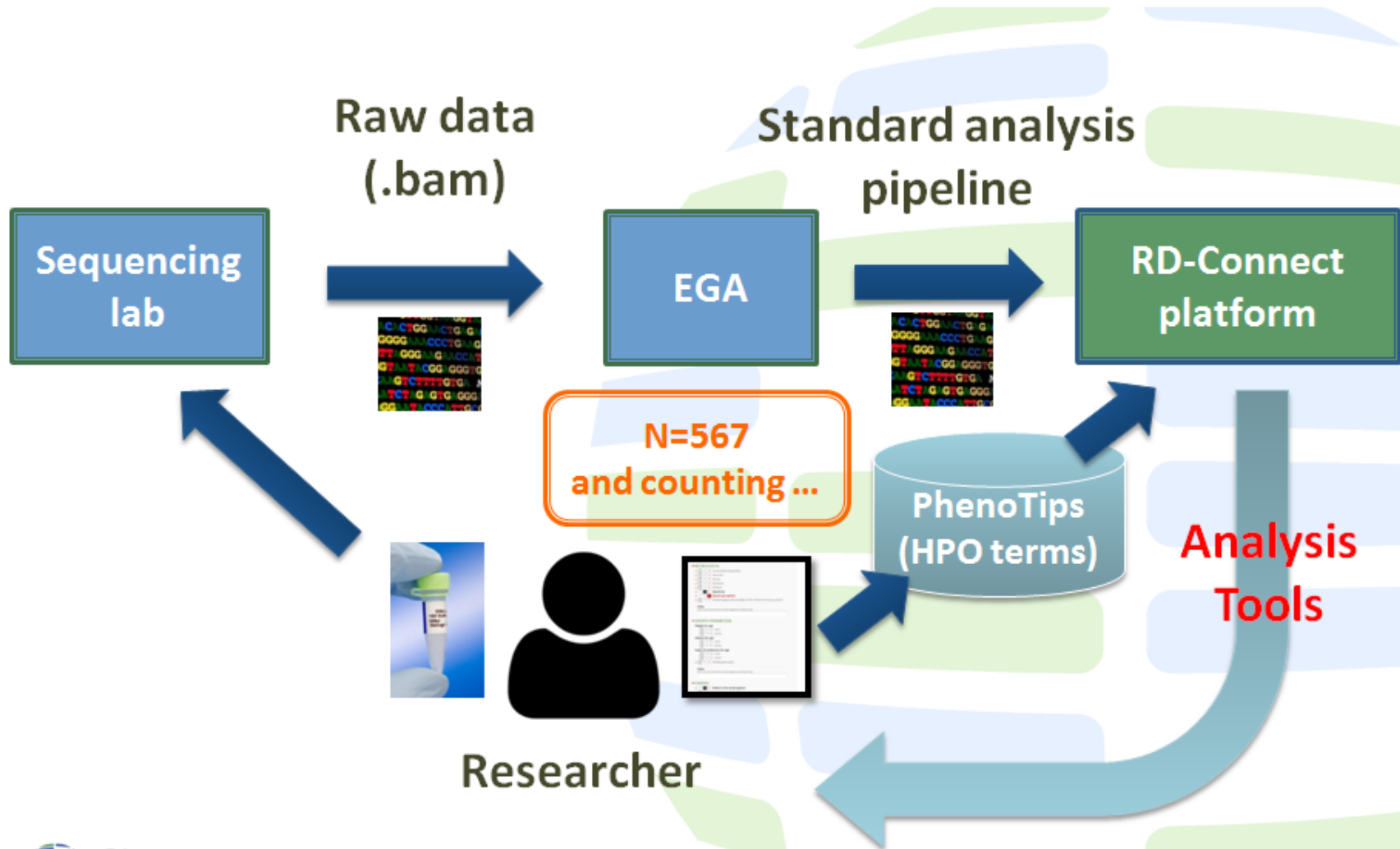
2





Data flow to RD-Connect

5



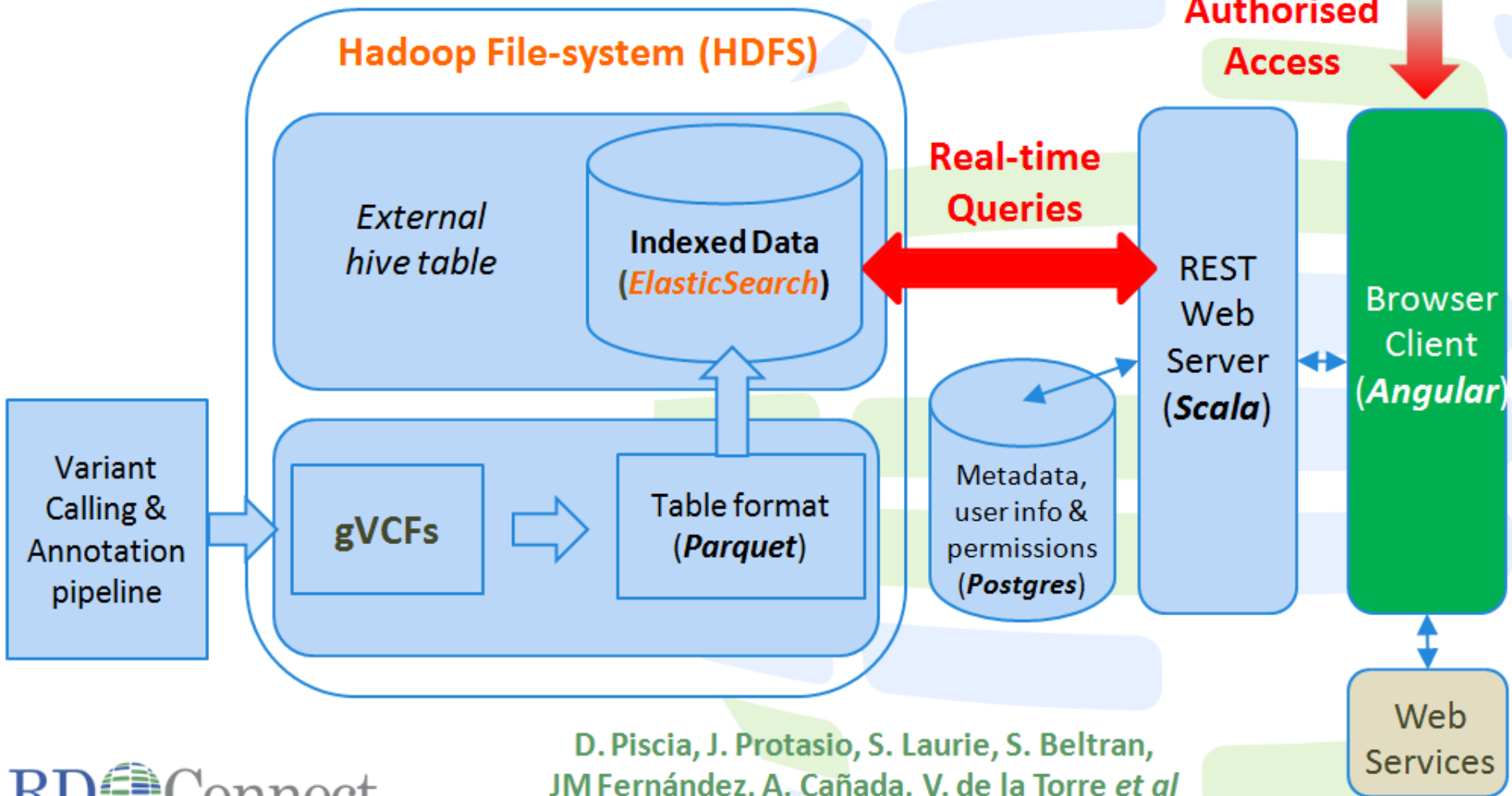


Genomics platform architecture

9



Authorised
Access



D. Piscia, J. Protasio, S. Laurie, S. Beltran,
JM Fernández, A. Cañada, V. de la Torre et al



RD-Connect Genomics Platform

Search Samples

LOG OUT



Genomics

RESET RUN QUERY

Variant Type: coding high moderate Population: gp1_af exac SNV->MT: A D SNV->SIFT: D SNV->PP2: D P

- Sample selection ? >
- Variant Type ? >
- Population ? >
- SNV Effect Prediction ? >
- Gene and Chromosome Coordinates >

Chrom	Pos	Ref	Alt
1	17302199	T	G

Functional Predictive Population Samples ALFA Diseasecard

Gene Name	Gene BioType	Transcript ID	Transcript BioType	Effect	Effect Impact	Function Class
MFAP2	CODING	ENST00000375535	protein_coding	NON_SYNONYMOUS_CODING	MODERATE	MISSENSE
MFAP2	CODING	ENST00000375534	protein_coding	NON_SYNONYMOUS_CODING	MODERATE	MISSENSE
MFAP2	CODING	ENST00000438542	protein_coding	NON_SYNONYMOUS_CODING	MODERATE	MISSENSE

Results 5 EXPORT ALL

First Previous 1 Next Last

Samples

Chrom	Pos	dbSNP	Ref	Alt	GT0000010	GT0000004	GT0000017	INDEL	Gene Name	Effect Impact	CADD	SIFT	PP2	MT	ExAC	1000GP AF
1	17302199		T	G	T/G	T/T	T/T		MFAP2	MODERATE	21.1	D	P	D	1.0E-4	0
13	77835380		G	A	G/A	G/G	G/G		MYCBP2	MODERATE	24.7	T	D	D	0	0
17	9066234		A	C	A/C	A/A	A/A		NTN1	MODERATE	27.5	D	P	D	0.0017	0
17	41584471		G	C	G/C	G/G	G/G		DHX8	MODERATE	16.3	T	B	D	0	0
19	39062811		G	C	G/C	G/G	G/G		RYR1	MODERATE	16.8	D	D	D	0	0



RD-Connect Genomics Platform

15

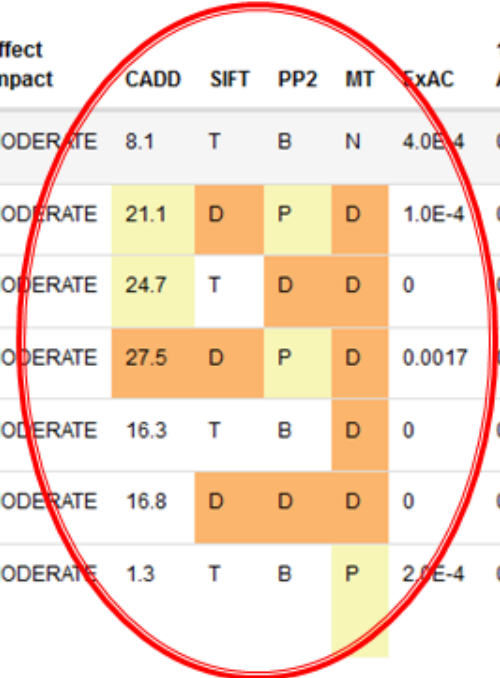
Results 7 [EXPORT ALL](#)

First Previous **1** Next Last

Variants

Color coding of pathogenicity predictors

Chrom	Pos	dbSNP	Ref	Alt	Candidate ?	GT ^{E000010}	GT ^{E000036}	GT ^{E000037}	INDEL	Gene Name	Effect Impact	CADD	SIFT	PP2	MT	ExAC	1000GP AF
1	12855648	rs200816125	A	G	0 <input type="button" value="add"/>	A/G	A/A	A/A		PRAMEF1	MODERATE	8.1	T	B	N	4.0E-4	0
1	17302199		T	G	0 <input type="button" value="add"/>	T/G	T/T	T/T		MFAP2	MODERATE	21.1	D	P	D	1.0E-4	0
13	77835365		G	A	0 <input type="button" value="add"/>	G/A	G/G	G/G		MYCBP2	MODERATE	24.7	T	D	D	0	0
17	9066234		A	C	0 <input type="button" value="add"/>	A/C	A/A	A/A		NTN1	MODERATE	27.5	D	P	D	0.0017	0
17	41584471		G	C	0 <input type="button" value="add"/>	G/C	G/G	G/G		DHX8	MODERATE	16.3	T	B	D	0	0
19	39062815		G	C	0 <input type="button" value="add"/>	G/C	G/G	G/G		RYR1	MODERATE	16.8	D	D	D	0	0
19	54746382	rs201396172	C	T	0 <input type="button" value="add"/>	C/T	C/C	C/C		LILRA6 LILRB3	MODERATE	1.3	T	B	P	2.0E-4	0





RD-Connect Genomics Platform

18

Results 7 EXPORT ALL

First Previous 1 Next Last

Variants

Variants tagged in RD-Connect

Chrom	Pos	dbSNP	Ref	Alt	Candidate ?	GT ^{E000010}	GT ^{E000036}	GT ^{E000037}	INDEL	Gene Name	Effect Impact	CADD	SIFT	PP2	MT	ExAC	1000GP AF
1	12855648	rs200816125	A	G	0 <input type="button" value="add"/>	A/G	A/A	A/A		PRAMEF1	MODERATE	8.1	T	B	N	4.0E-4	0
1	17302199		T	G	0 <input type="button" value="add"/>	T/G	T/T	T/T		MFAP2	MODERATE	21.1	D	P	D	1.0E-4	0
13	77835365		G	A	0 <input type="button" value="add"/>	G/A	G/G	G/G		MYCBP2	MODERATE	24.7	T	D	D	0	0
17	9066234		A	C	3 <input type="button" value="add"/>					NTN1	MODERATE	27.5	D	P	D	0.0017	0
17	41584471		G	C	0 <input type="button" value="add"/>					DHX8	MODERATE	16.3	T	B	D	0	0
19	39062815		G	C	0 <input type="button" value="add"/>					RYR1	MODERATE	16.8	D	D	D	0	0
19	54746382	rs201396172	C	T	0 <input type="button" value="add"/>	C/A				LILRA6 LILRB3	MODERATE	1.3	T	B	P	2.0E-4	0

this variant has been tagged 3 time(s) at:
 E000361 with significance pathogenic
 E000002 with significance benign
 E000001 with significance pathogenic



RD-Connect Genomics Platform

19

Results 7 EXPORT ALL

Variants

Tag a variant

Chrom	Pos	dbSNP	Ref	Alt	Candidate ?	
1	12855648	rs200816125	A	G	0	<input type="button" value="add"/>
1	17302199		T	G	0	<input type="button" value="add"/>
13	77835365		G	A	0	<input type="button" value="add"/>
17	9066234		A	C	0	<input type="button" value="add"/>
17	41584471		G	C	0	<input type="button" value="add"/>
19	39062815		G	C	0	<input type="button" value="add"/>
19	54746382	rs201396172	C	T	0	<input type="button" value="add"/>

Identifying variant as causal

Researcher username
j.protasio

Date
07/03/2016

Sample
E000002

Mode of inheritance
x-linked dominant

Origin
somatic

Clinical significance
likely pathogenic

Pubmed IDs
199514

Comments (evidence and/or experiments done and any other relevant details)

ClinVar categories

	ADD	SIFT	PP2	MT	ExAC	1000GP AF
1	T	B	N	4.0E-4	0	
1	D	P	D	1.0E-4	0	
7	T	D	D	0	0	
5	D	P	D	0.0017	0	
3	T	B	D	0	0	
8	D	D	D	0	0	
8	T	B	P	2.0E-4	0	

Protasio, S. Laurie, A. Papakonstantinou, S. Beltran



GA4GH Beacon Project

24

A **beacon** is a simple web service that answers questions:

Question: Is this variant present in your dataset?

Answer: Yes/No

<http://ga4gh.org/#/beacon>

Posted: May 29, 2015

Beacon Project

Being implemented on the website of the world's top genomic organizations to test the willingness of international sites to share genetic data.



About this Project

The **Beacon project** is a project to test the willingness of international sites to share genetic data in the simplest of all technical contexts. It is defined as a simple public web service that any institution can implement as a service. The service is designed merely to accept a query of the form "Do you have any genomes with an 'A' at position 100,735 on chromosome 3" (or similar data) and responds with one of "Yes" or "No." A site offering this service is called a "beacon". This open web service is designed to be technically simple, easy to implement, and to not return privacy violating information.



GA4GH Beacon Network

25

https://beacon-network.org//#/

GRCh37 ▾

13 : 32954208 A>T

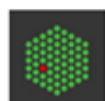
Search

Response All None

- Found 8
- Not Found 43
- Error 8

Organization All None

- AMPLab, University of C...
- BGI
- BioReference Laboratories
- Broad Institute
- Centre for Genomic Regu...
- CNAG
- Curoverse
- DNASTack
- EMBL European Bioinfor...
- Global Alliance for Geno...
- Google
- Institute for Systems Biol...
- Mike Lin
- National Center for Biote...



EBI - 1000 Genomes Project, ...

EMBL European Bioinformatics Institute

Not Found



ExAC

Broad Institute

Not Found



ICGC - Cancer Projects

Ontario Institute for Cancer Research

Not Found



Kaviar

Institute for Systems Biology

Found



NHLBI Exome Sequence Proj...

National Center for Biotechnology Information

Not Found



RD-Connect

CNAG

Not Found



Not Found



The MatchMaker Exchange (MME, IRDiRC, GA4GH)

26

<http://www.matchmakerexchange.org/>



HOW TO GET STARTED

OUR RESOURCE LIBRARY

EXCHANGE PARTICIPANTS

CONTACT US

Matchmaker Exchange

Genomic discovery through the exchange of phenotypic & genotypic profiles



- Data sits isolated in databases
- Sometimes difficult to find another case with a variant in the same gene
- MME is working towards a federated platform to facilitate matching of cases with similar phenotypic and genotypic profiles.



The MatchMaker Exchange (MME, IRDiRC, GA4GH)



Question: Do you have a patient with similar phenotype and genotype as mine?

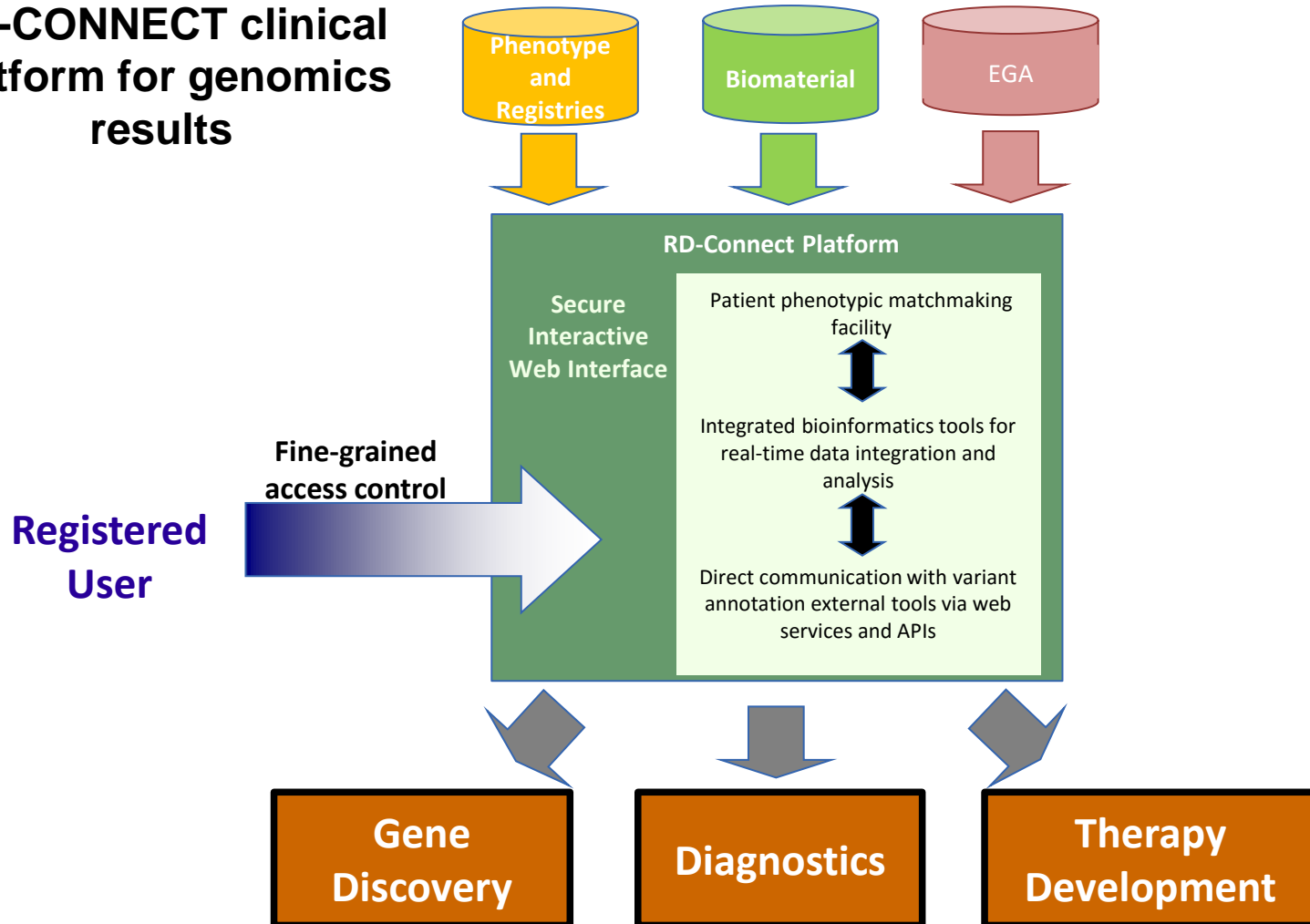
- Two-sided matching
- One-sided matching

Human Mutation

Volume 36, Issue 10, pages 915-921, 17 SEP 2015 DOI: 10.1002/humu.22858 <http://onlinelibrary.wiley.com/doi/10.1002/humu.22858/full#humu22858-fig-0001>

Personalized Medicine

RD-CONNECT clinical platform for genomics results





Data in the RD-Connect platform

NeurOOmics
1000 exomes

EURenOOmics
1000 exomes

MyoSeq, Newcastle
1000 exomes

SeqNMD, Broad Institute
500 exomes

RD  **Connect**

NCNP, Japan
500 exomes

CIMG, Slovenia
300 exomes

CNAG, Spain
300 exomes



WE WANT YOU

Final Remarks

- Importance of standards
- Provenance of data is important
- With assured quality there is no need for storing raw data on high end, spinning disks. Tape storage is cheaper and saves on electricity.
- Primary analysis close to large volume data – bandwidth
- Who is the data user? Clinician (competence – web-service or command line?)
- Data sharing increases benefit to patient
- Rare disease is a use case for omics data in other diseases



cnag

baldiri reixac, 4
pcb - tower i, 2nd floor
08028 barcelona

t +34 93 4020542
f +34 93 4037279
www.cnag.eu



cnag