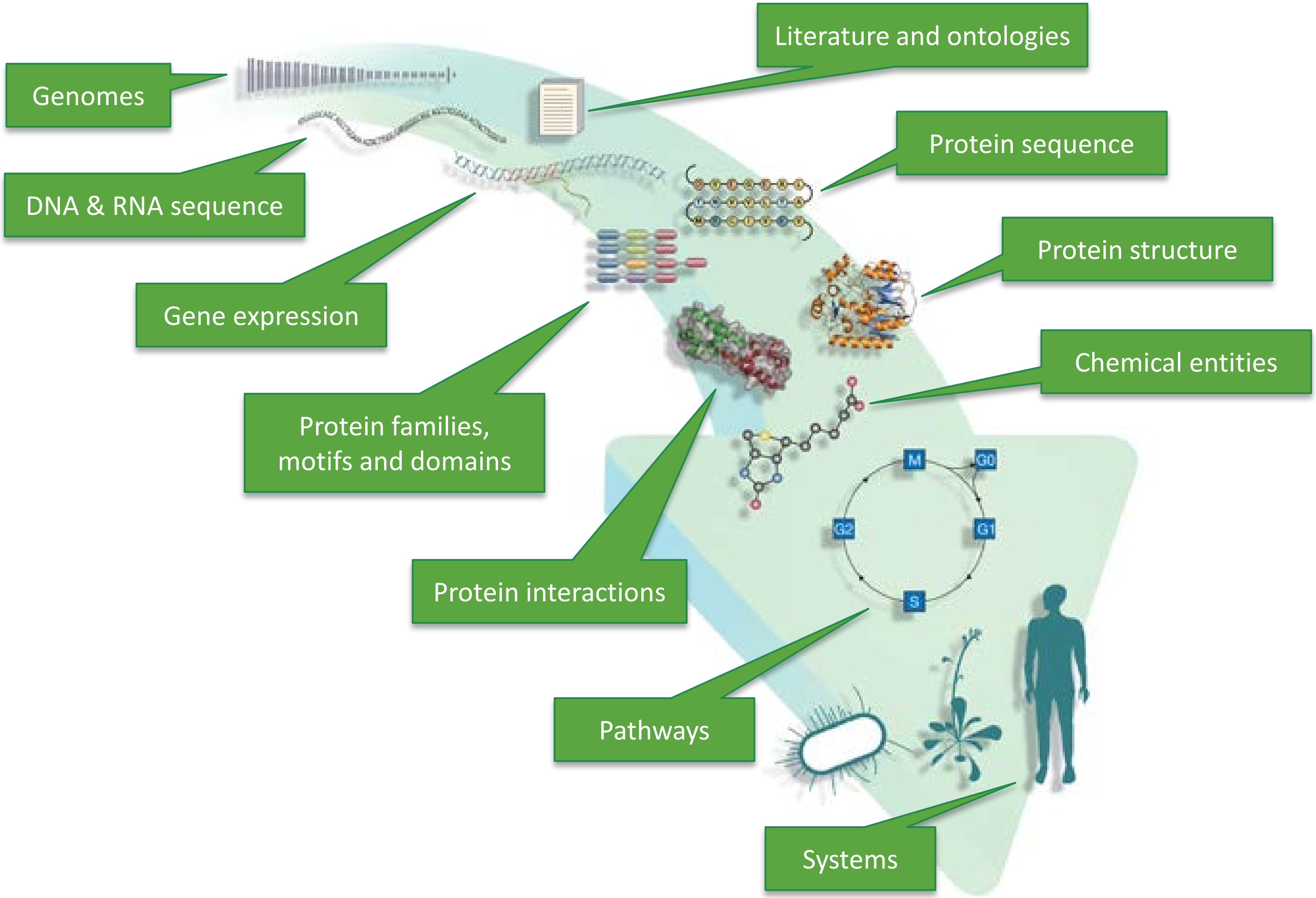
'Bioinformatics in academia as related to eHealth' - including the "Genomic Virtual Lab"

Dr Gareth Price

Head of Computational Biology

Queensland Facility of Advanced Bioinformatics

From Genomes to Systems



Genomes

Literature and ontologies

DNA & RNA sequence

Protein sequence

Gene expression

Protein structure

Chemical entities

Protein families, motifs and domains

Protein interactions

Pathways

Systems

EMBL Australia | Bioinformatics Resource

What is Next Gen Sequencing?

Next Generation Sequencing:

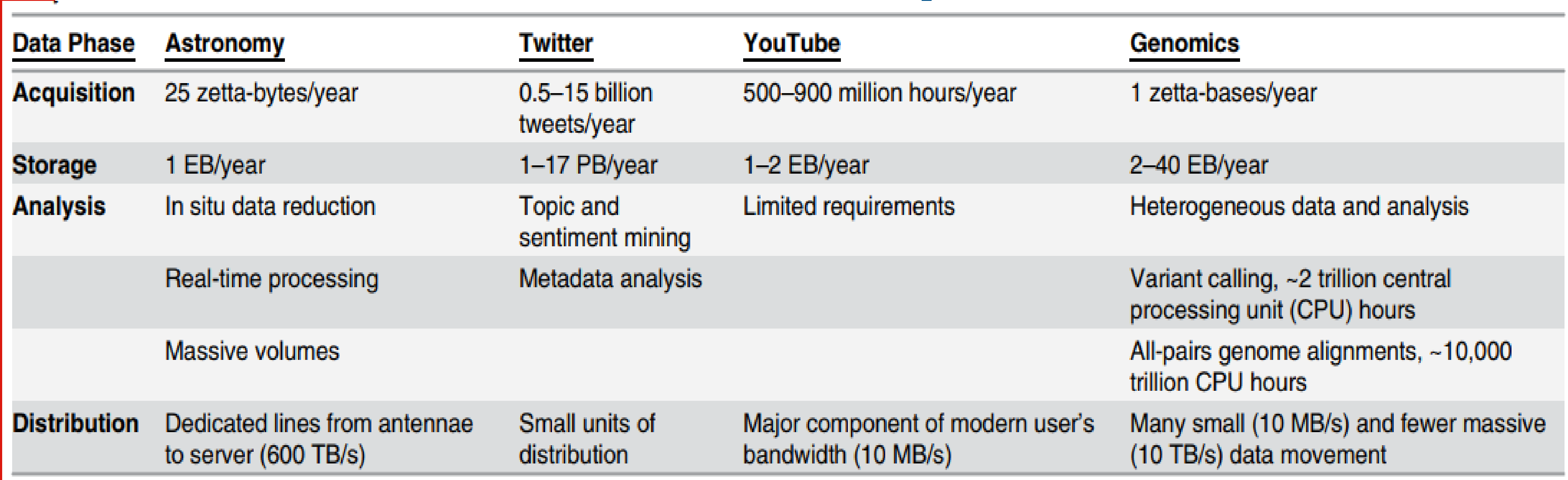
- high-throughput sequencing
- massive parallel short (and now long) read sequencing
- deep sequencing

In reality NGS really just refers to the scale of sequencing. For Example:

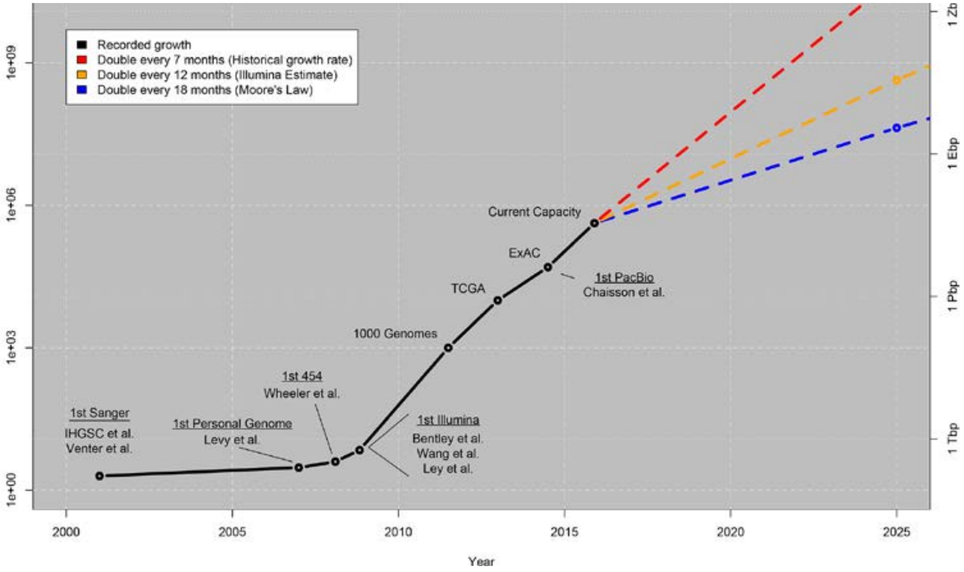
ABI Sequencers, Venter Institute – 2007
1 Human Genome in total (years!)

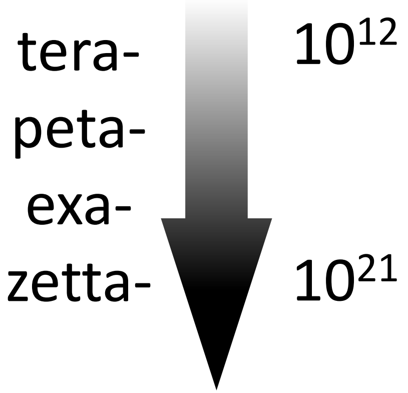
Illumina HiSeq 2000s, BGI – 2013
2 Human Genomes per machine (days!)

Big Data: Acquisition, Storage, Distribution, and Analysis



Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement





tera-

peta-

exa-

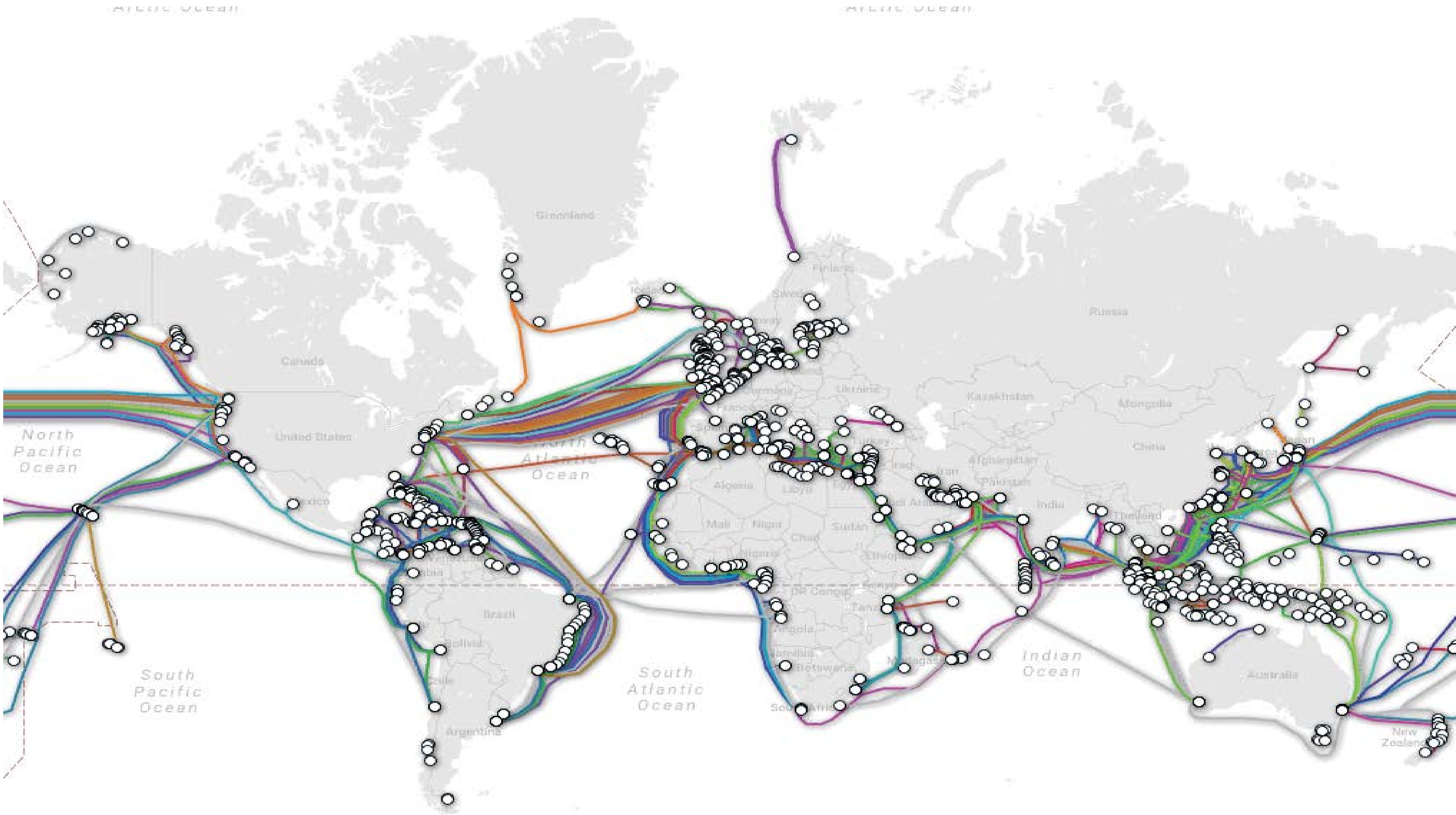
zetta-

10^{12}

10^{21}

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C et al., PLOS Biology (2015)

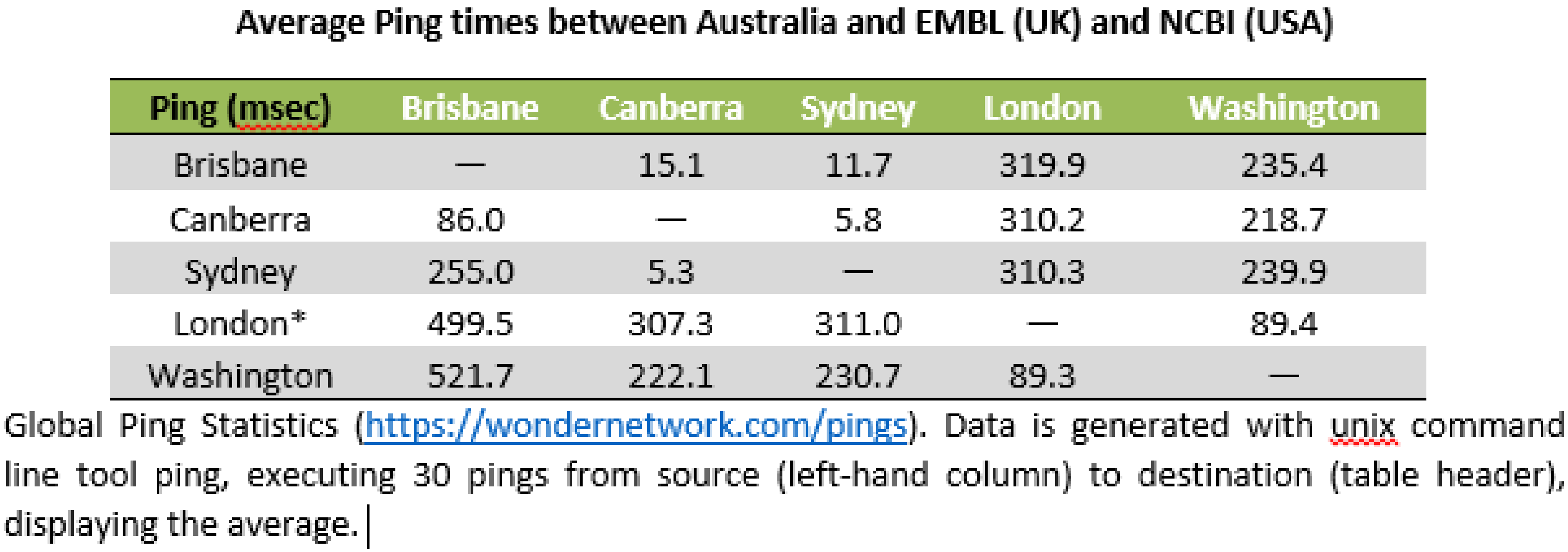
Submarine Cable Map TeleGeography



www.submarinecablemap.com

Data Transfer speeds

- Datasets potentially very large (many Tbs)
- Download time is lengthy and at risk of failure and interruption

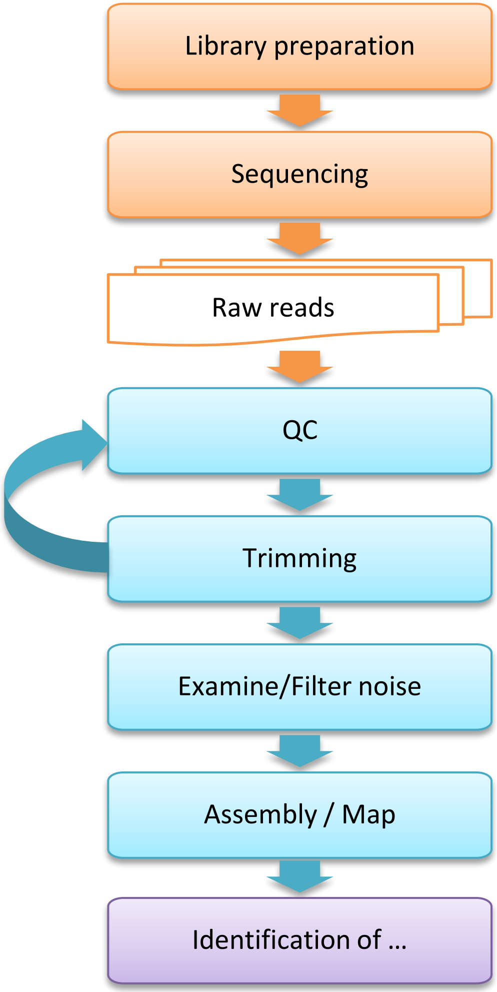


Average Ping times between Australia and EMBL (UK) and NCBI (USA)

Ping (msec)	Brisbane	Canberra	Sydney	London	Washington
Brisbane	—	15.1	11.7	319.9	235.4
Canberra	86.0	—	5.8	310.2	218.7
Sydney	255.0	5.3	—	310.3	239.9
London*	499.5	307.3	311.0	—	89.4
Washington	521.7	222.1	230.7	89.3	—

Global Ping Statistics (https://wondernetwork.com/pings). Data is generated with unix command line tool ping, executing 30 pings from source (left-hand column) to destination (table header), displaying the average.

Overview of NGS data flow



Library preparation

preparation of samples in the lab

Sequencing

library run in sequencer either at sequencing facility or in-house

Raw reads

sequencer generates output "**reads**"

QC

quality checks of the reads

Trimming

trimming reads if required

Examine/Filter noise

Removal of foreign DNA materials

Assembly / Map

Construction of genomes using sequencing reads or count transcripts

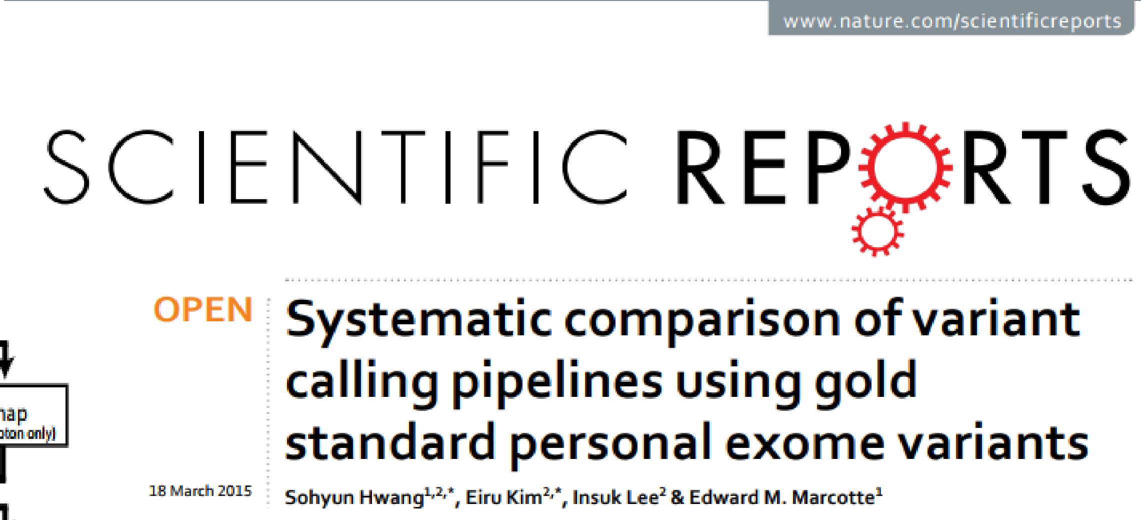
Identification of …

Branches out to research analysis

Tools – Academic and Clinical

- **Freeware**
  - **Genome Analysis Toolkit (GATK)**
  - **Virtual Labs / Machines**
    - Galaxy
    - R Studio (Bioconductor)
    - Command Line
- **Commercial**
  - **Agilent**
    - Cartagenia Bench Lab for Molecular Pathology
  - **Illumina**
    - BaseSpace
  - **Qiagen**
    - CLC-Bio Suite of Analysis Products
    - Ingenuity Pathway Analysis
    - Ingenuity Variant Analysis
    - ANNOVAR
  - **ThermoFisher**
    - Ion Reporter
  - *Google Genomics*
  - *Microsoft Genomics*
  - *Oracle Healthcare Precision Medicine*

  http://grouthbio.com/Genome_Software_Service.php



www.nature.com/scientificreports

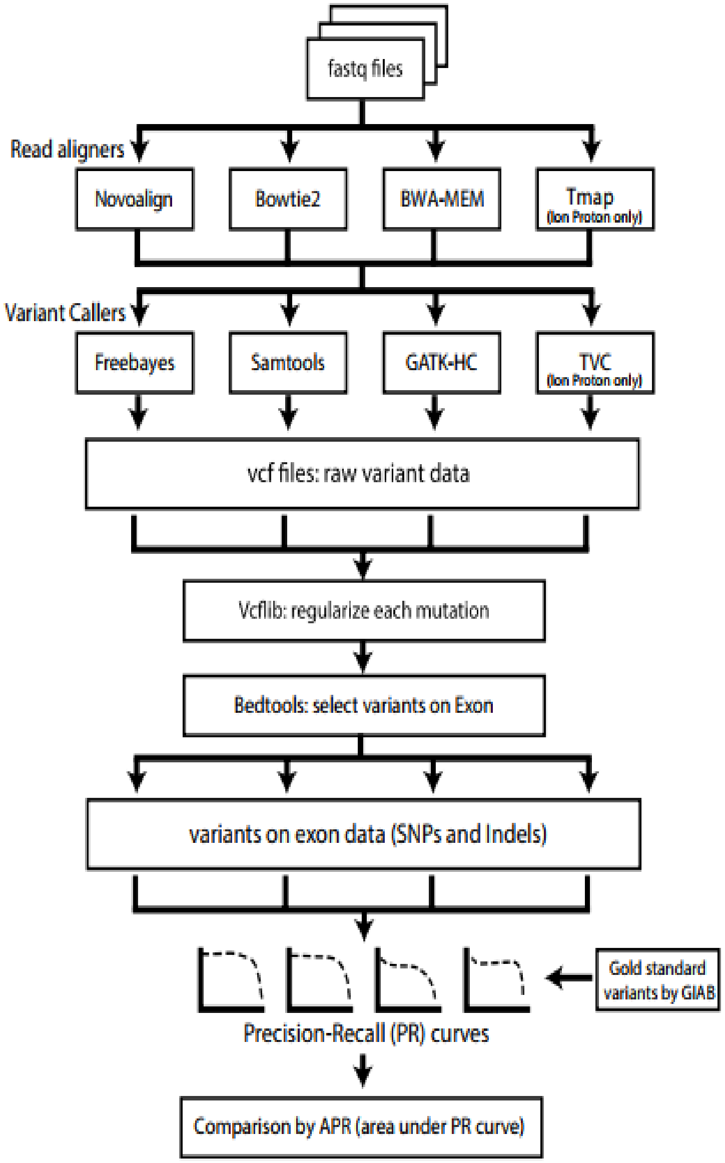
SCIENTIFIC REPORTS

OPEN

18 March 2015

Systematic comparison of variant calling pipelines using gold standard personal exome variants

Sohyun Hwang[1,2,*], Eiru Kim[2,*], Insuk Lee[2] & Edward M. Marcotte[1]



fastq files

Read aligners

Novoalign

Bowtie2

BWA-MEM

Tmap (Ion Proton only)

Variant Callers

Freebayes

Samtools

GATK-HC

TVC (Ion Proton only)

vcf files: raw variant data

Vcflib: regularize each mutation

Bedtools: select variants on Exon

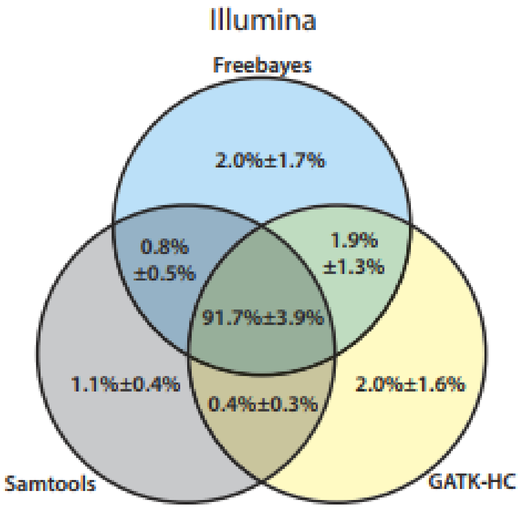
variants on exon data (SNPs and Indels)

Gold standard variants by GIAB

Precision-Recall (PR) curves

Comparison by APR (area under PR curve)

“We observed different biases toward specific types of SNP genotyping errors by the different variant callers”



Health Solutions to the use Genomic Information

- American College of Medical Genetics
  - 2015: doi:10.1038/gim.2015.30
  - *Benign, Likely Benign, Of Unknown Significance, Likely Pathogenic and Pathogenic*

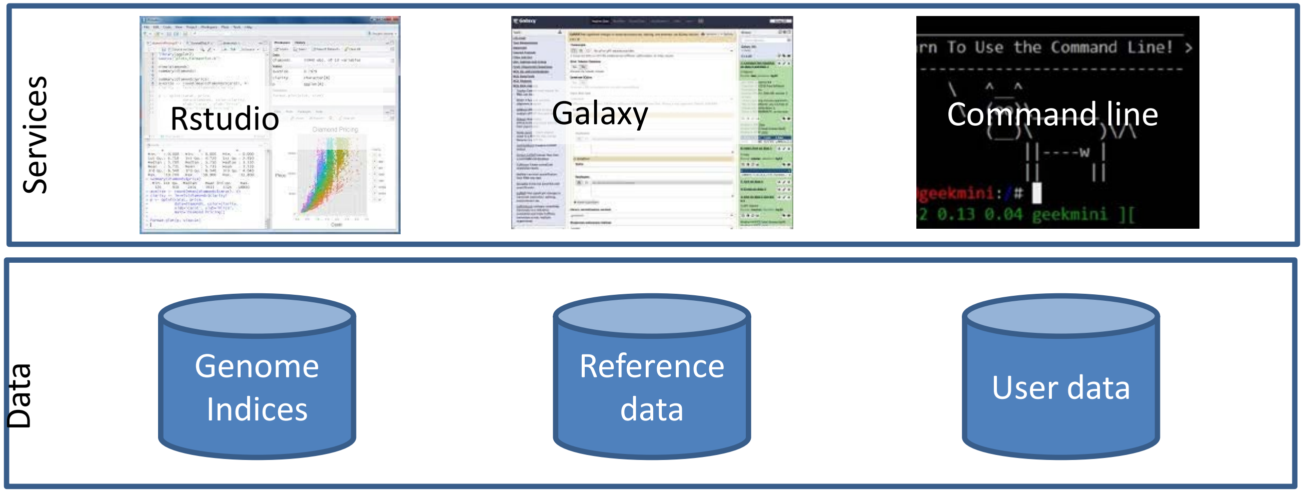
- NIH – National Human Genome Research Institute: Division of Genomic Medicine
  - Undiagnosed Rare Disorders, GWAS studies, report formats

- Global Alliance for Genomics and Health (GA4GH)
  - policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework

Genomics Virtual Labs

- Virtual Machines with pre-installed suite of tools for performing bioinformatics analyses
- Public instances all around the world
- gvl.org.au
- usegalaxy.org.au [currently Galaxy Qld]
- GVL provides compute and storage
- Galaxy houses reference data, public data and indices

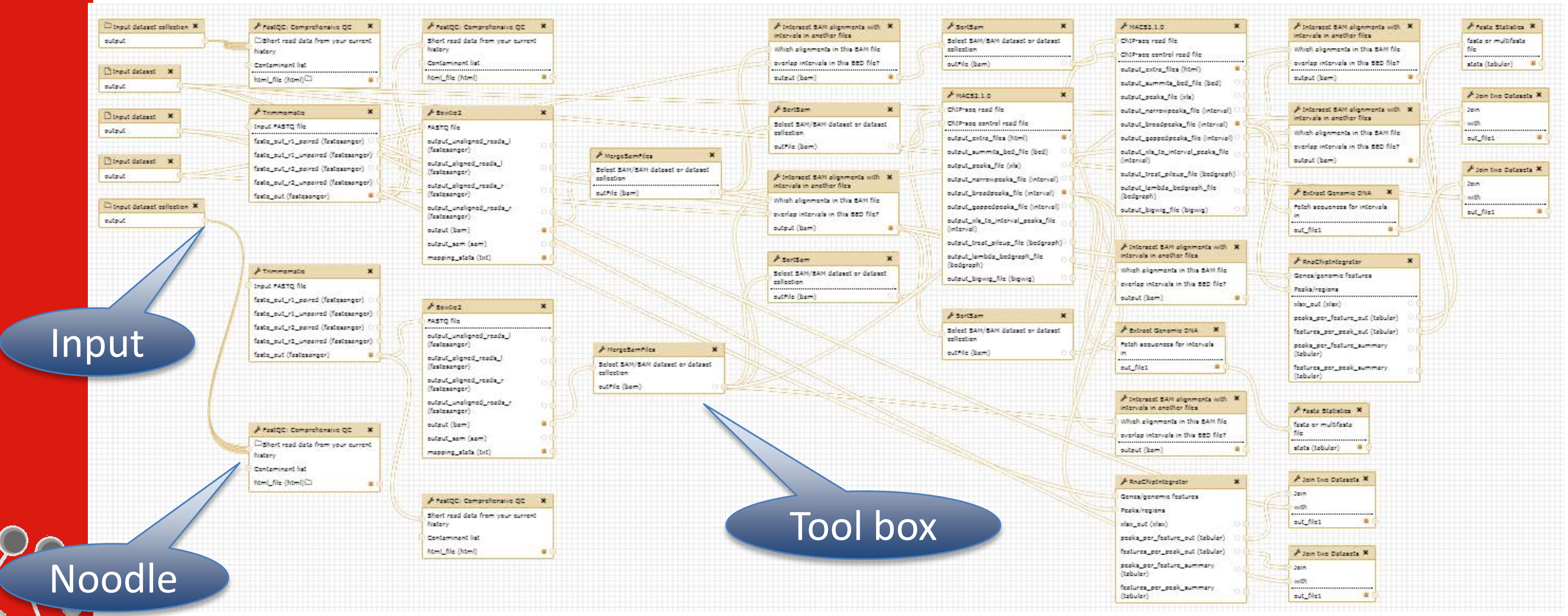
- Galaxy houses your uploaded data, computed data and results
- Direct import from public repositories
- Data visualisation options
- Data sharing (with and without data duplication)
- Big community
- Easy registration



nectar THE UNIVERSITY OF MELBOURNE THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Galaxy is a workflow engine

A Galaxy workflow is a series of tools and dataset actions that run in sequence as a batch operation

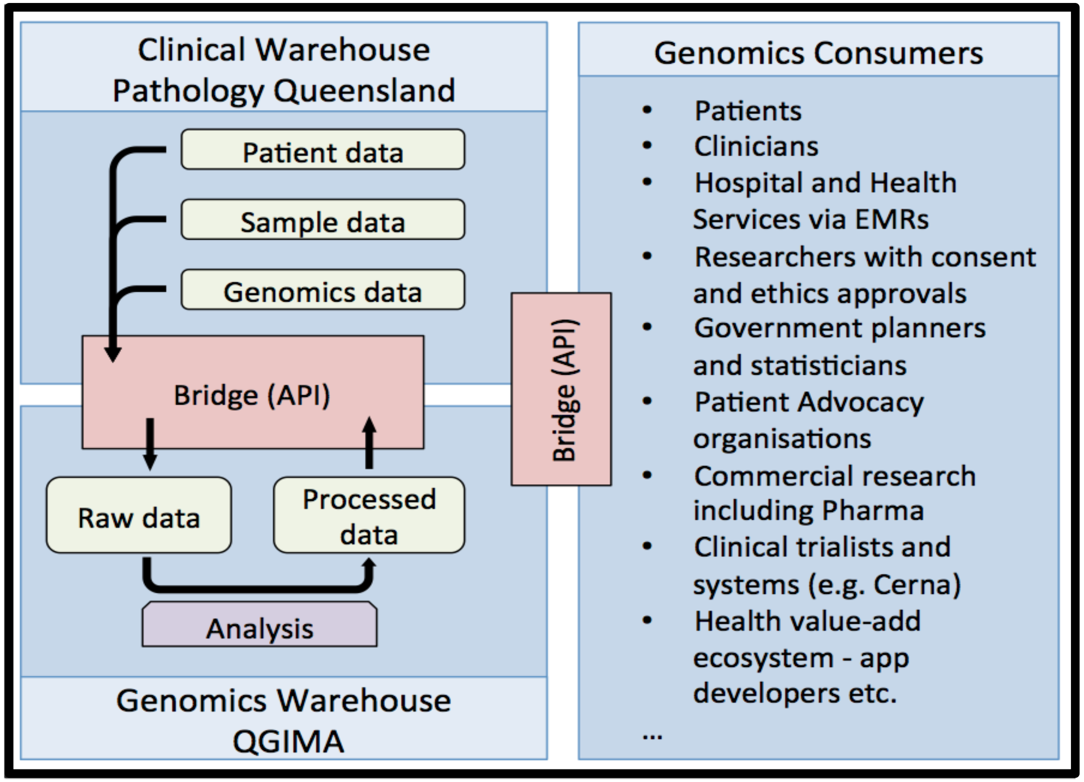

Input dataset collection
output
Input dataset
output
Input dataset
output
Input dataset
output
Input dataset collection
output
FastQC: Comprehensive QC
Short read data from your current history
Contaminant list
html_file (html)
Trimmomatic
Input FASTQ file
fastq_out_r1_paired (fastqsanger)
fastq_out_r1_unpaired (fastqsanger)
fastq_out_r2_paired (fastqsanger)
fastq_out_r2_unpaired (fastqsanger)
fastq_out (fastqsanger)
Trimmomatic
Input FASTQ file
fastq_out_r1_paired (fastqsanger)
fastq_out_r1_unpaired (fastqsanger)
fastq_out_r2_paired (fastqsanger)
fastq_out_r2_unpaired (fastqsanger)
fastq_out (fastqsanger)
FastQC: Comprehensive QC
Short read data from your current history
Contaminant list
html_file (html)
FastQC: Comprehensive QC
Short read data from your current history
Contaminant list
html_file (html)
Bowtie2
FASTQ file
output_unaligned_reads_l (fastqsanger)
output_aligned_reads_l (fastqsanger)
output_aligned_reads_r (fastqsanger)
output_unaligned_reads_r (fastqsanger)
output (bam)
output_sam (sam)
mapping_stats (txt)
Bowtie2
FASTQ file
output_unaligned_reads_l (fastqsanger)
output_aligned_reads_l (fastqsanger)
output_aligned_reads_r (fastqsanger)
output_unaligned_reads_r (fastqsanger)
output (bam)
output_sam (sam)
mapping_stats (txt)
FastQC: Comprehensive QC
Short read data from your current history
Contaminant list
html_file (html)
MergeSamFiles
Select SAM/BAM dataset or dataset collection
outFile (bam)
MergeSamFiles
Select SAM/BAM dataset or dataset collection
outFile (bam)
Intersect BAM alignments with intervals in another files
Which alignments in this BAM file
overlap intervals in this BED file?
output (bam)
SortSam
Select SAM/BAM dataset or dataset collection
outFile (bam)
Intersect BAM alignments with intervals in another files
Which alignments in this BAM file
overlap intervals in this BED file?
output (bam)
SortSam
Select SAM/BAM dataset or dataset collection
outFile (bam)
SortSam
Select SAM/BAM dataset or dataset collection
outFile (bam)
MACS2.1.0
ChIP-seq read file
ChIP-seq control read file
output_extra_files (html)
output_summits_bed_file (bed)
output_peaks_file (xls)
output_narrowpeaks_file (interval)
output_broadpeaks_file (interval)
output_gappedpeaks_file (interval)
output_xls_to_interval_peaks_file (interval)
output_treat_pileup_file (bedgraph)
output_lambda_bedgraph_file (bedgraph)
output_bigwig_file (bigwig)
SortSam
Select SAM/BAM dataset or dataset collection
outFile (bam)
MACS2.1.0
ChIP-seq read file
ChIP-seq control read file
output_extra_files (html)
output_summits_bed_file (bed)
output_peaks_file (xls)
output_narrowpeaks_file (interval)
output_broadpeaks_file (interval)
output_gappedpeaks_file (interval)
output_xls_to_interval_peaks_file (interval)
output_treat_pileup_file (bedgraph)
output_lambda_bedgraph_file (bedgraph)
output_bigwig_file (bigwig)
Intersect BAM alignments with intervals in another files
Which alignments in this BAM file
overlap intervals in this BED file?
output (bam)
Extract Genomic DNA
Fetch sequences for intervals in
out_file1
Intersect BAM alignments with intervals in another files
Which alignments in this BAM file
overlap intervals in this BED file?
output (bam)
RnaChipIntegrator
Genes/genomic features
Peaks/regions
xlsx_out (xlsx)
peaks_per_feature_out (tabular)
features_per_peak_out (tabular)
peaks_per_feature_summary (tabular)
features_per_peak_summary (tabular)
Intersect BAM alignments with intervals in another files
Which alignments in this BAM file
overlap intervals in this BED file?
output (bam)
Intersect BAM alignments with intervals in another files
Which alignments in this BAM file
overlap intervals in this BED file?
output (bam)
Extract Genomic DNA
Fetch sequences for intervals in
out_file1
RnaChipIntegrator
Genes/genomic features
Peaks/regions
xlsx_out (xlsx)
peaks_per_feature_out (tabular)
features_per_peak_out (tabular)
peaks_per_feature_summary (tabular)
features_per_peak_summary (tabular)
Fasta Statistics
fasta or multifasta file
stats (tabular)
Join two Datasets
Join
with
out_file1
Join two Datasets
Join
with
out_file1
Join two Datasets
Join
with
out_file1
Join two Datasets
Join
with
out_file1
Fasta Statistics
fasta or multifasta file
stats (tabular)
Input
Noodle
Tool box

Genomic Information Management

Queensland Genomics Health Alliance

ADVANCE QUEENSLAND | Queensland Government

- *Improving the health of Queenslanders by delivering genomic medicine*



Clinical Warehouse
Pathology Queensland

Patient data

Sample data

Genomics data

Bridge (API)

Raw data

Processed data

Analysis

Genomics Warehouse
QGIMA

Bridge (API)

Genomics Consumers

- Patients
- Clinicians
- Hospital and Health Services via EMRs
- Researchers with consent and ethics approvals
- Government planners and statisticians
- Patient Advocacy organisations
- Commercial research including Pharma
- Clinical trialists and systems (e.g. Cerna)
- Health value-add ecosystem - app developers etc.

…

Leads

David Hansen, AeHRC, CSIRO
John Pearson, QIMRB MRI

Collaborators

Dominique Gorse, QCIF
Paul Leo, QUT
Pamela Pollock, QUT
Cas Simons, UQ
Nic Waddell, QIMRB MRI
Amanda Spurdle, QIMRB MRI
Sunil Lakhani, PQ & UQ
Naomi Wray, IMB, QBI & UQ
Hugo Leroux, AeHRC, CSIRO
Alejandro Metke, AeHRC, CSIRO

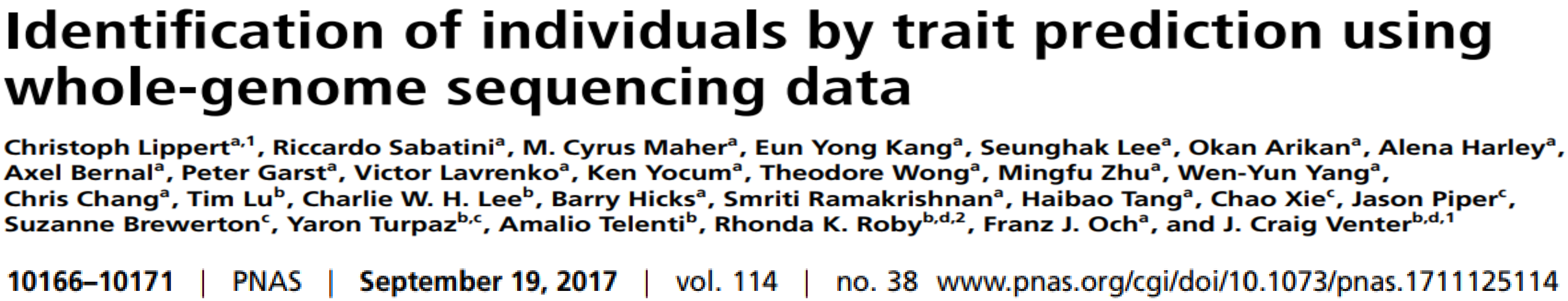
Contacts

www.qfab.org

contact@qfab.org
training@qfab.org

g.price@qfab.org

Balancing Sharing with Identification

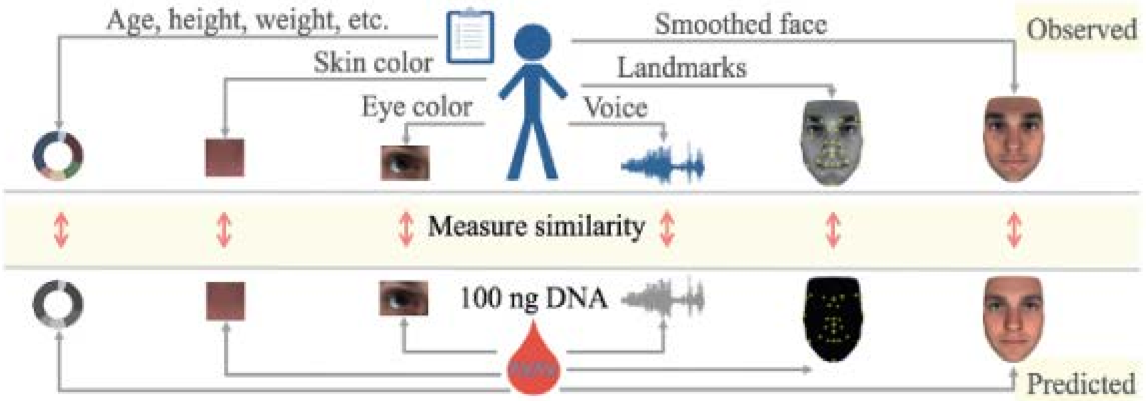


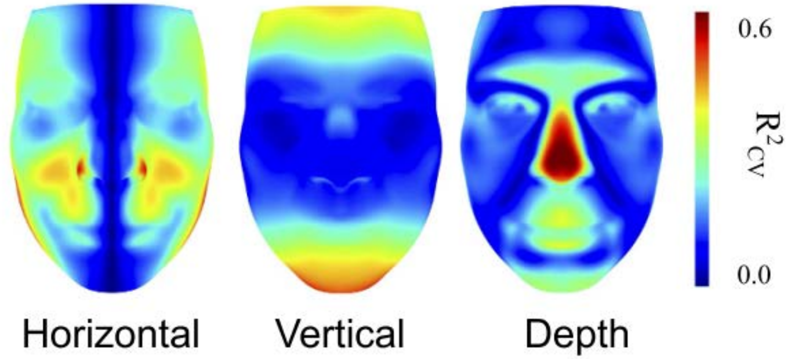
Identification of individuals by trait prediction using whole-genome sequencing data

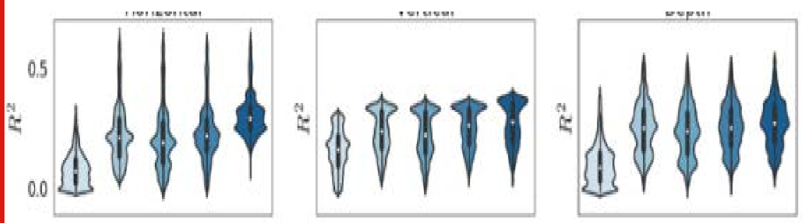
Christoph Lippert[a,1], Riccardo Sabatini[a], M. Cyrus Maher[a], Eun Yong Kang[a], Seunghak Lee[a], Okan Arikan[a], Alena Harley[a], Axel Bernal[a], Peter Garst[a], Victor Lavrenko[a], Ken Yocum[a], Theodore Wong[a], Mingfu Zhu[a], Wen-Yun Yang[a], Chris Chang[a], Tim Lu[b], Charlie W. H. Lee[b], Barry Hicks[a], Smriti Ramakrishnan[a], Haibao Tang[a], Chao Xie[c], Jason Piper[c], Suzanne Brewerton[c], Yaron Turpaz[b,c], Amalio Telenti[b], Rhonda K. Roby[b,d,2], Franz J. Och[a], and J. Craig Venter[b,d,1]

10166–10171 | PNAS | September 19, 2017 | vol. 114 | no. 38 www.pnas.org/cgi/doi/10.1073/pnas.1711125114

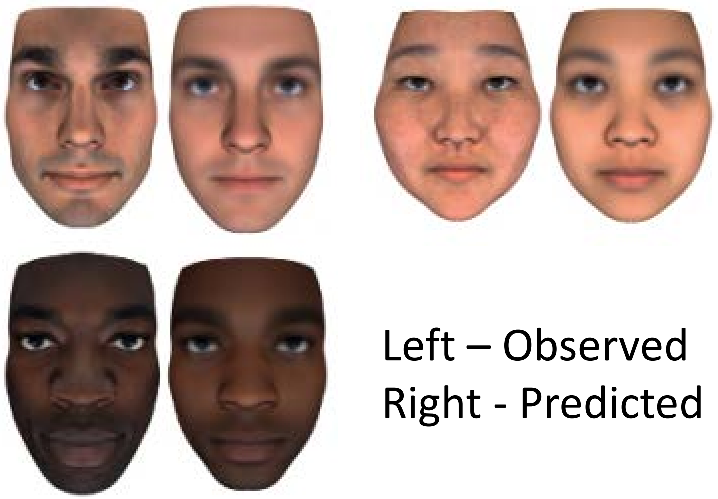
- 3D facial structure
- Voice
- Biological age
- Height
- Weight
- BMI
- Eye colour
- Skin colour
- Sex
- *Hair Colour*
- *Baldness*







Sex + Ancestry + SNPs + Age + BMI



Left – Observed
Right - Predicted

- limitations in statistical power (n=1,061)
- individually, each model provided limited information about an individual's identity
- multiple prediction models enabled matching between genomes and phenotypic profiles with good accuracy
- "Over time, predictions will get more precise…
- … thus, the results of this work will be of greater consideration in the current discussion on genome privacy protection."

- **How we protect against identification?**
  – Mask known predictor sites?