TGAC

The Genome Analysis Centre

Building Excellence in Genomics and Computational Bioscience

The Genome Analysis Centre

Big Data & Algorithmics in Biotechnology

Mario Caccamo

mario.caccamo@tgac.ac.uk
@mcaccamo

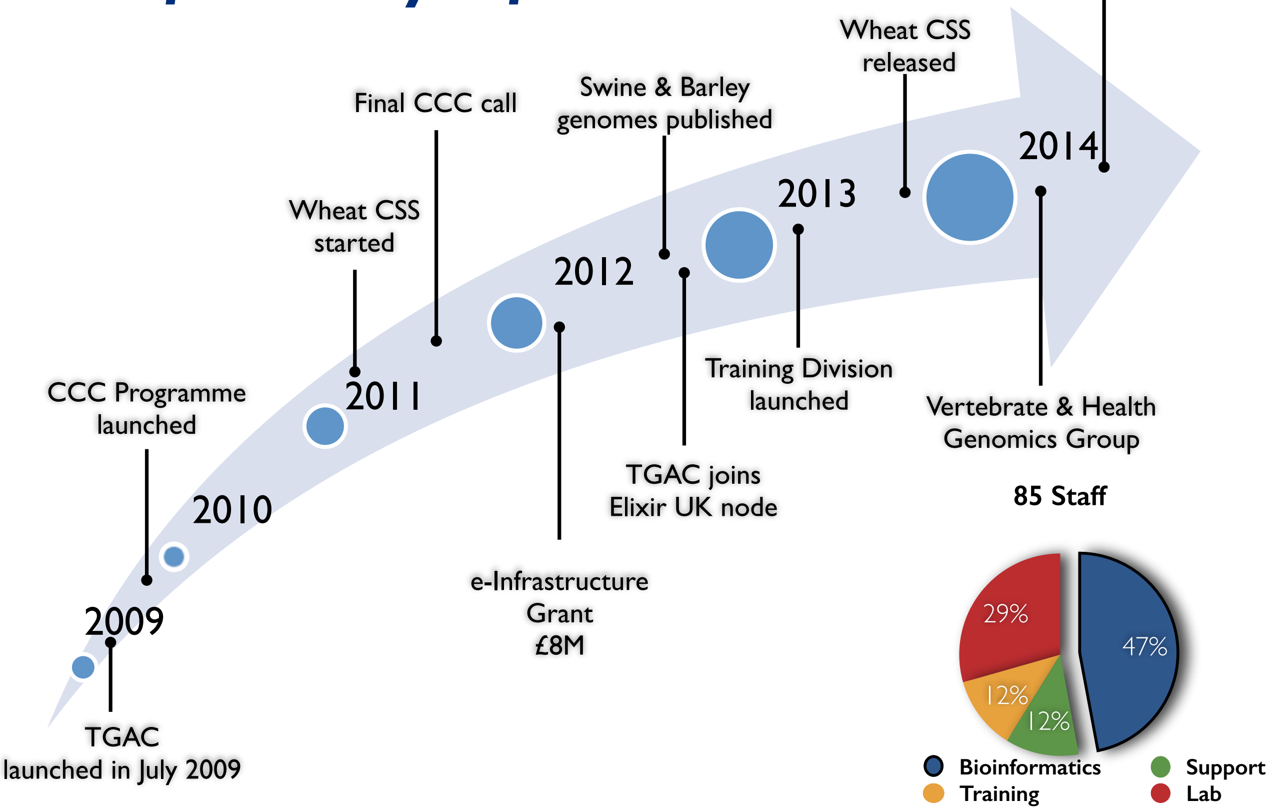
TGAC
The Genome Analysis Centre™

BBSRC

Greater Norwich
Development
Partnership

A brief history of TGAC

DNA Synthesis Grant £1.9M



2009

TGAC launched in July 2009

CCC Programme launched

2010

Wheat CSS started

2011

Final CCC call

e-Infrastructure Grant £8M

2012

Swine & Barley genomes published

TGAC joins Elixir UK node

Training Division launched

2013

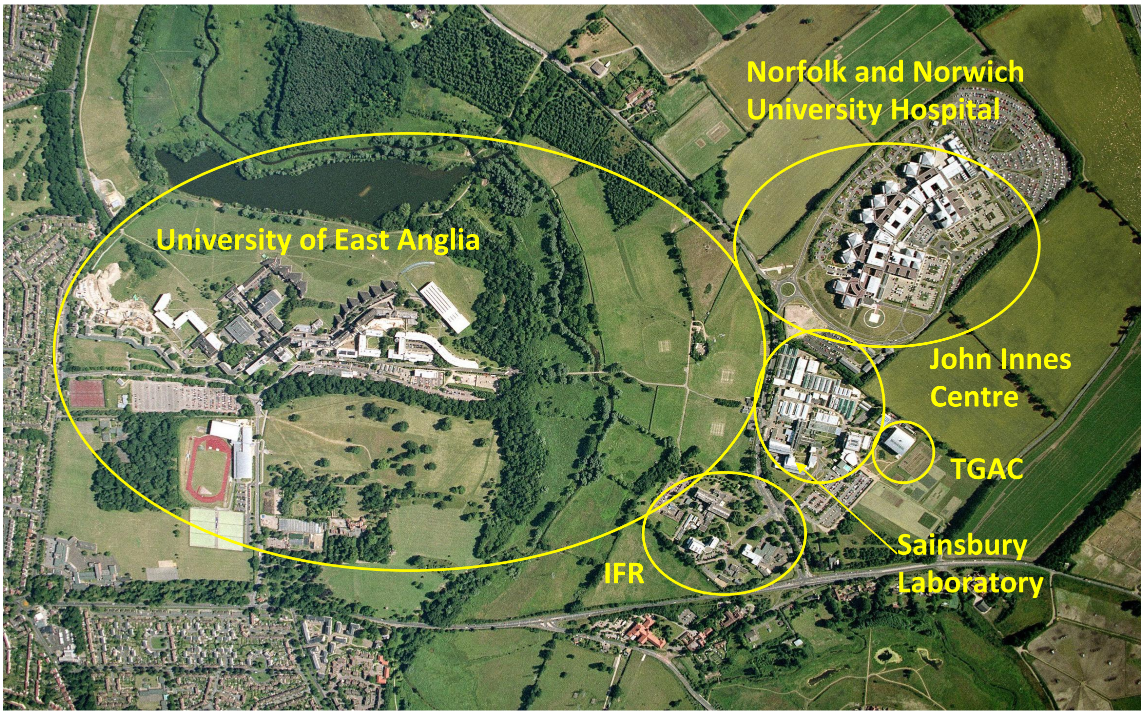
Wheat CSS released

2014

Vertebrate & Health Genomics Group

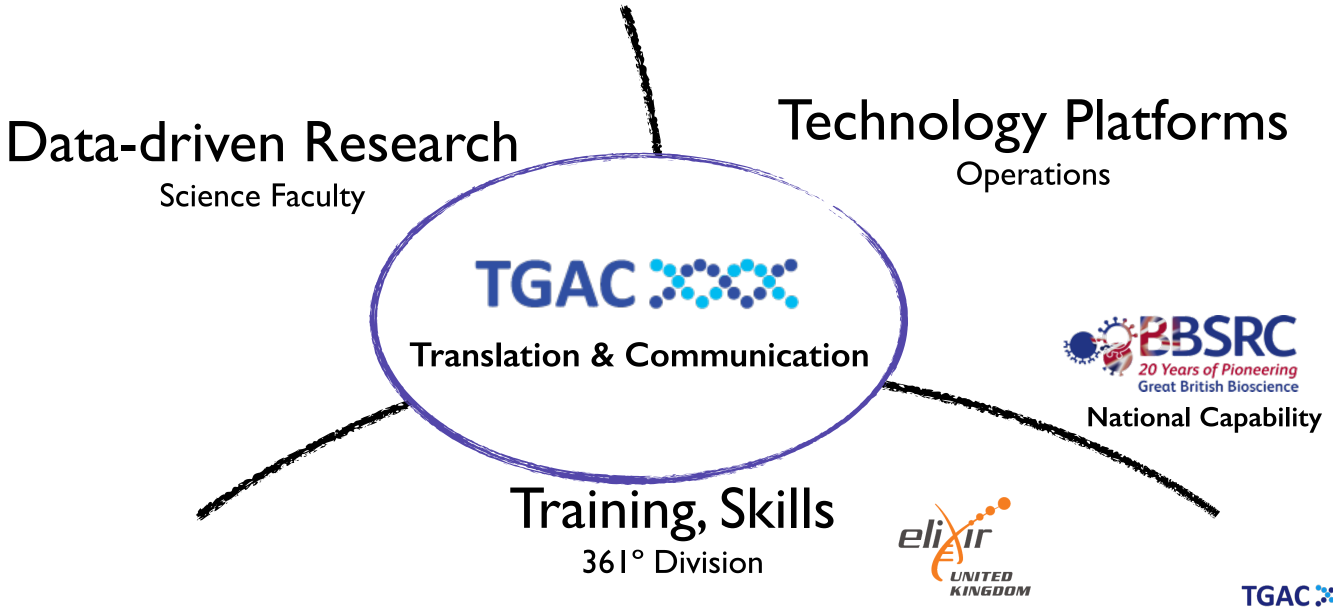
85 Staff

Norwich Research Park


Norfolk and Norwich University Hospital
University of East Anglia
John Innes Centre
TGAC
IFR
Sainsbury Laboratory

Vision

“To be a centre of excellence in ***high-end*** **genomics** and **computational biology** to develop the Norwich Research Park as world leaders in bioinformatics and biotechnology.”



Data-driven Research

Science Faculty

Technology Platforms

Operations

Translation & Communication

National Capability

Training, Skills

361° Division

Technology Platforms

Illumina HiSeq

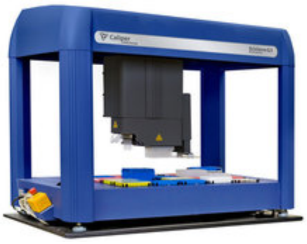
~300 Gbps/day

(100 times human genome)

Pacific Bioscience RS II
single-molecule sequencing
1of 3 in the UK

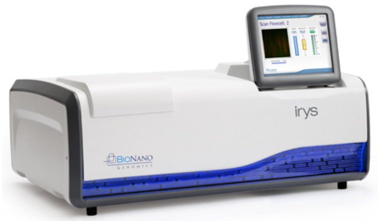
SGI UVs

2500 cores
20 Tbytes RAM

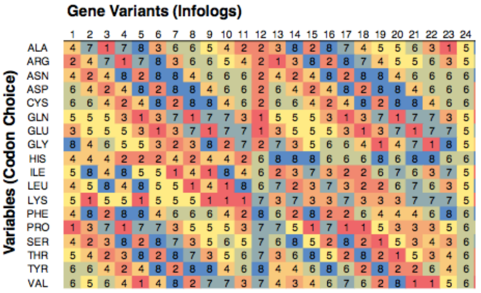


Lab Automation
PerkinElmer Reference Lab

New Technologies



BioNano Irys
(Optical Mapping)



Gene Variants (Infologs)

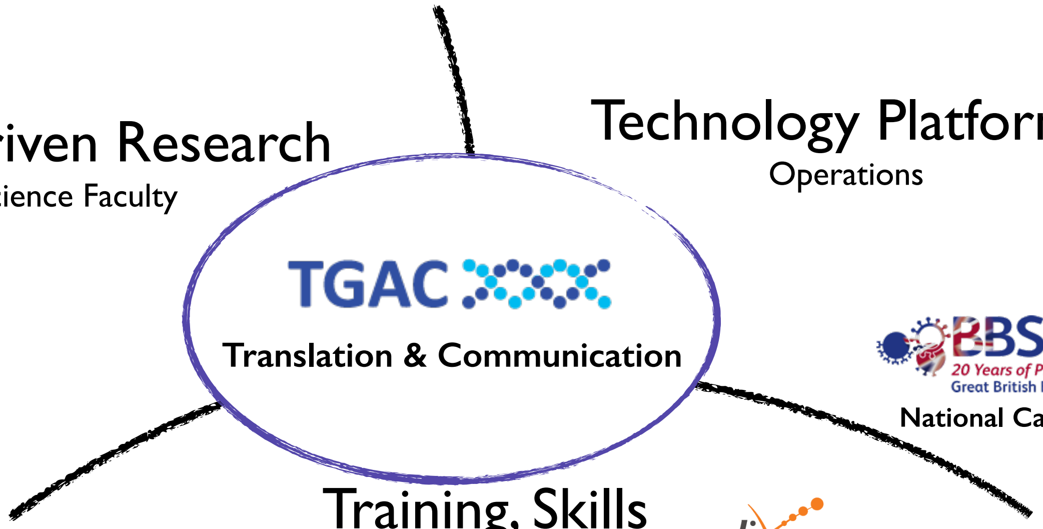
Variables (Codon Choice)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
ALA	4	7	1	7	8	3	6	6	5	4	2	2	3	8	2	8	7	4	5	5	6	3	1	5
ARG	2	4	7	1	7	8	3	6	6	5	4	2	1	3	8	2	8	7	4	5	5	6	3	5
ASN	4	2	4	8	2	8	8	4	6	6	6	2	4	2	4	8	2	8	8	4	6	6	6	6
ASP	6	4	2	4	8	2	8	8	4	6	6	2	6	4	2	4	8	2	8	8	4	6	6	6
CYS	6	6	4	2	4	8	2	8	8	4	6	2	6	6	4	2	4	8	2	8	8	4	6	6
GLN	5	5	5	3	1	3	7	1	7	7	3	1	5	5	5	3	1	3	7	1	7	7	3	5
GLU	3	5	5	5	3	1	3	7	1	7	7	1	3	5	5	5	3	1	3	7	1	7	7	5
GLY	8	4	6	5	5	3	2	3	8	2	7	2	7	3	5	6	6	4	1	4	7	1	8	5
HIS	4	4	4	2	2	2	4	2	4	4	2	6	8	8	8	6	6	6	8	6	8	8	6	6
ILE	5	8	4	8	5	5	1	4	1	8	4	6	2	3	7	3	2	2	6	7	6	3	7	5
LEU	4	5	8	4	8	5	5	1	4	1	8	6	7	2	3	7	3	2	2	6	7	6	3	5
LYS	5	1	5	5	1	5	5	5	1	1	1	7	3	7	3	3	7	3	3	3	7	7	7	5
PHE	4	8	2	8	8	4	6	6	6	4	2	8	6	2	8	2	2	6	4	4	4	6	8	6
PRO	1	3	7	1	7	7	3	5	5	5	3	7	7	5	1	7	1	1	5	3	3	3	5	6
SER	4	2	3	8	2	8	7	3	5	6	5	7	6	8	5	2	8	2	1	5	3	4	3	6
THR	5	4	2	3	8	2	8	7	3	5	6	7	3	6	8	5	2	8	2	1	5	3	4	6
TYR	6	6	4	2	4	8	2	8	8	4	6	8	4	4	6	8	6	2	8	2	2	6	4	6
VAL	6	5	6	4	1	4	8	2	7	7	3	7	4	3	4	6	7	6	2	8	1	1	5	6

DNA Synthesis Grant
£1.9M

MinIon Nanopore Sequencing
(Oxford Nanopore Technologies)

Vision

"To be a centre of excellence in ***high-end* genomics** and **computational biology** to develop the Norwich Research Park as world leaders in bioinformatics and biotechnology."



Data-driven Research

Science Faculty

Technology Platforms

Operations

TGAC

Translation & Communication

BBSRC 20 Years of Pioneering Great British Bioscience

National Capability

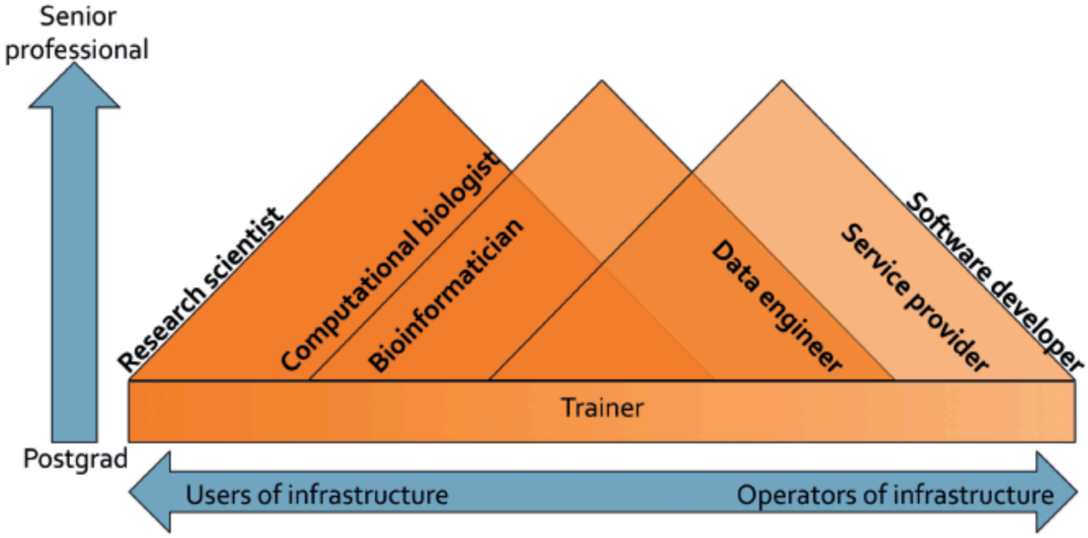
Training, Skills

361° Division

elixir UNITED KINGDOM

Elixir UK

Bioinformatics Training!

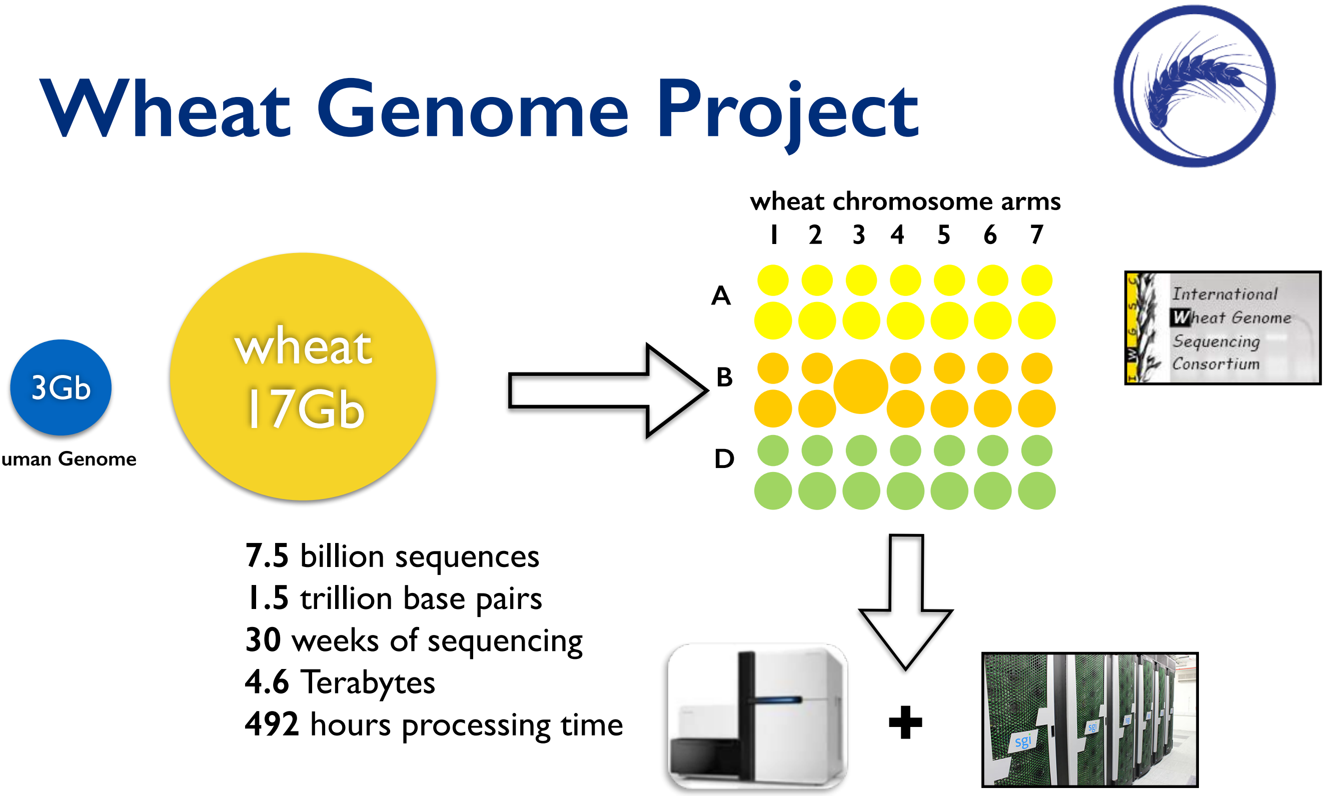


TransPlant

WP12 - Implementation of resource-intensive algorithms for plant genomics data.

- To evaluate the adequacy of the current **resource-intensive** algorithms for plant genomics data.
- To build a **network of developers** with expertise in algorithms for plant genomics.
- To implement **novel algorithmic** solutions for concrete challenges in large plant genomes.

Wheat Genome Project



7.5 billion sequences
1.5 trillion base pairs
30 weeks of sequencing
4.6 Terabytes
492 hours processing time

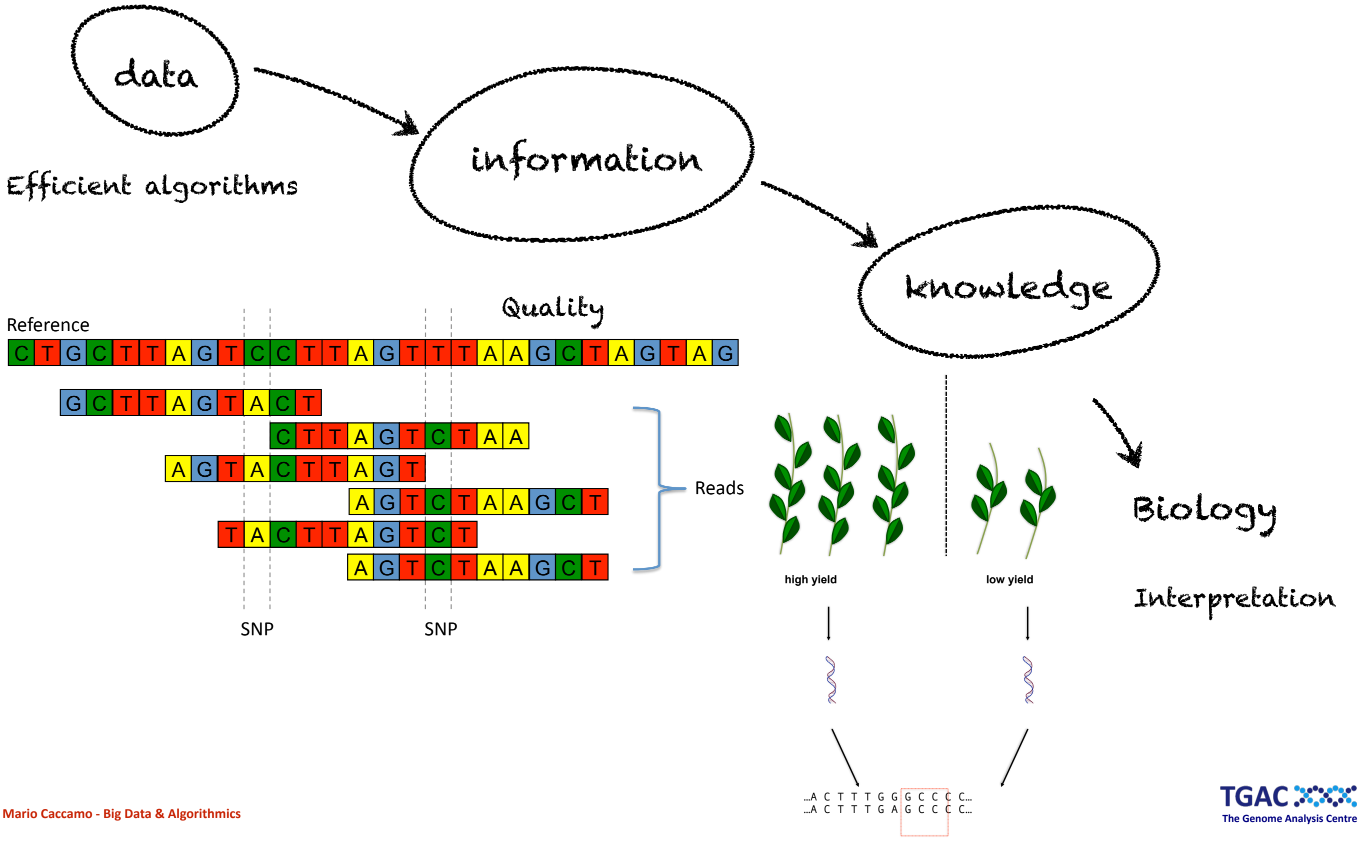
10M Sequences!

First wheat
whole genome sequence
released in September 2013

BBSRC sLOLa Grant
Triticeae Genomics
for sustainable agriculture
(TGAC, JIC, EBI, RRes)

Sarah Ayling
Matt Clark

From Genotype to Phenotype

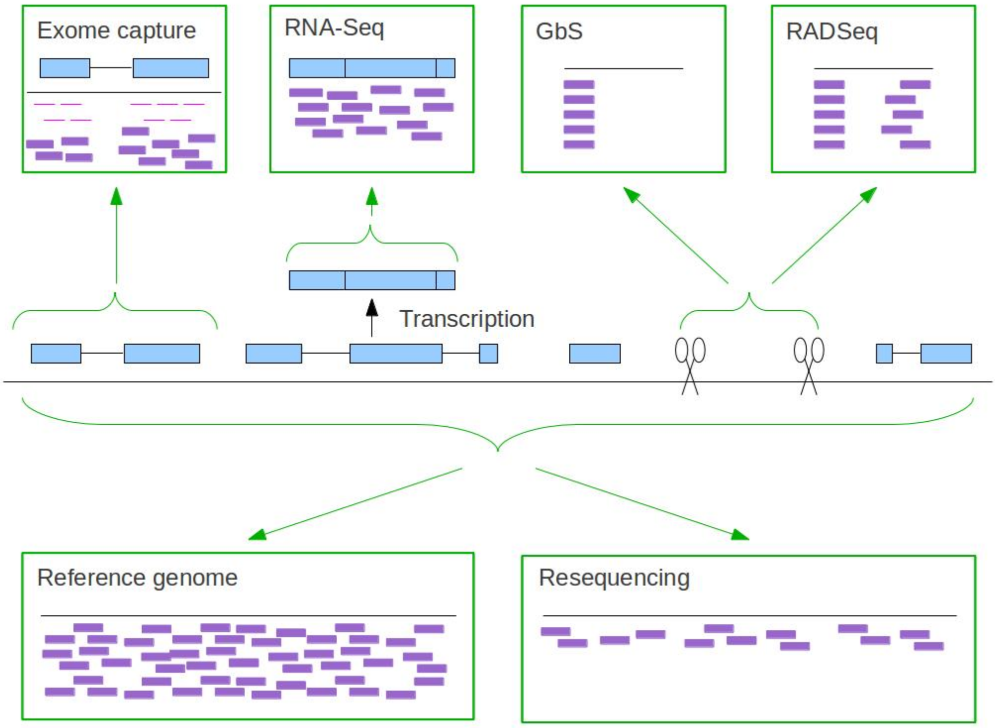


Reference
C T G C T T A G T C C T T A G T T T A A G C T A G T A G

G C T T A G T A C T
C T T A G T C T A A
A G T A C T T A G T
A G T C T A A G C T
T A C T T A G T C T
A G T C T A A G C T

…A C T T T G G G C C C C…
…A C T T T G A G C C C C…

Screening for Diversity



Exome Captures

Barley capture

61.6Mb capture:

- **150,000 Morex exons**
- **35,000 full length cDNAs**
- **110,000 RNAseq contigs**

Mascher *et al*., The Plant Journal (2013)

Wheat capture

84Mb capture:

- **57,000 *T. turgidum* RNAseq**
- **24,000 Public wheat cDNAs**
- **1,800 barley cDNAs**

Krasileva *et al*., Genome Biology 2013, 14:R66

Surveillance & Diagnostics

- **Technologies**
  - real-time data generation and analysis (Nanopore)
  - DNA extraction
- **Transmission Studies**
  - model zoonotic diseases
  - swine and horse flu transmission studies
- **Swine Flu Dynamics**
  - sLoLa (Pirbrigth, Cambridge)
- **Nornex (Ash Dieback)**
  - Genomics & genetics to tackle ash dieback
  - JIC, TSL (Norwich), Exeter, Edinburgh

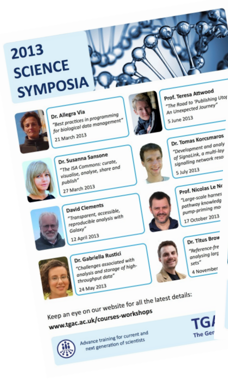
Bioinformatics Software

MISO
Cortex
TGAC Browser
RAMPART
KAT
NextClip
Bubbleparse
StatsDB
RADplex
kONTAMINANT
BioJS

Rob Davey

Bioinformatics Training

- **Launched March 2013**
  - ➡ 14 courses
  - ➡ 280 trainees
- **Elixir**
  - ➡ Elixir UK focused on Bioinformatics Training
- **Norwich Research Park**
  - ➡ Year in Industry
  - ➡ Immersive visitors training
  - ➡ TGAC Symposia



2013 SCIENCE SYMPOSIA

Dr. Allegra Via
"Best practices in programming for biological data management"
21 March 2013

Dr. Susanna Sansone
"The ISA Commons: curate, visualise, analyse, share and publish"
27 March 2013

David Clements
"Transparent, accessible, reproducible analysis with Galaxy"
12 April 2013

Dr. Gabriella Rustici
"Challenges associated with analysis and storage of high-throughput data"
24 May 2013

Prof. Teresa Attwood
"The Road to 'Publishing Utop
An Unexpected Journey"
5 June 2013

Dr. Tomas Korcsmaros
"Development and analy
of SignaLink, a multi-lay
signalling network reso
5 July 2013

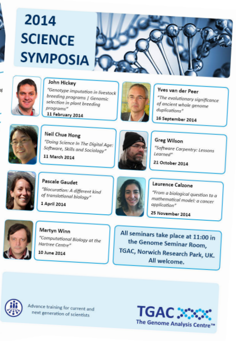
Prof. Nicolas Le No
"Large-scale harnes
pathway knowledg
pump-priming mo
17 October 2013

Dr. Titus Brow
"Reference-fre
analysing larg
sets"
4 November

Keep an eye on our website for all the latest details:
www.tgac.ac.uk/courses-workshops

Advance training for current and next generation of scientists

TGA
The Ger



2014 SCIENCE SYMPOSIA

John Hickey
"Genotype imputation in livestock breeding programs | Genomic selection in plant breeding programs"
11 February 2014

Yves van der Peer
"The evolutionary significance of ancient whole genome duplications"
16 September 2014

Neil Chue Hong
"Doing Science In The Digital Age: Software, Skills and Sociology"
11 March 2014

Greg Wilson
"Software Carpentry: Lessons Learned"
21 October 2014

Pascale Gaudet
"Biocuration: A different kind of translational biology"
1 April 2014

Laurence Calzone
"From a biological question to a mathematical model: a cancer application"
25 November 2014

Martyn Winn
"Computational Biology at the Hartree Centre"
10 June 2014

All seminars take place at 11:00 in the Genome Seminar Room, TGAC, Norwich Research Park, UK. All welcome.

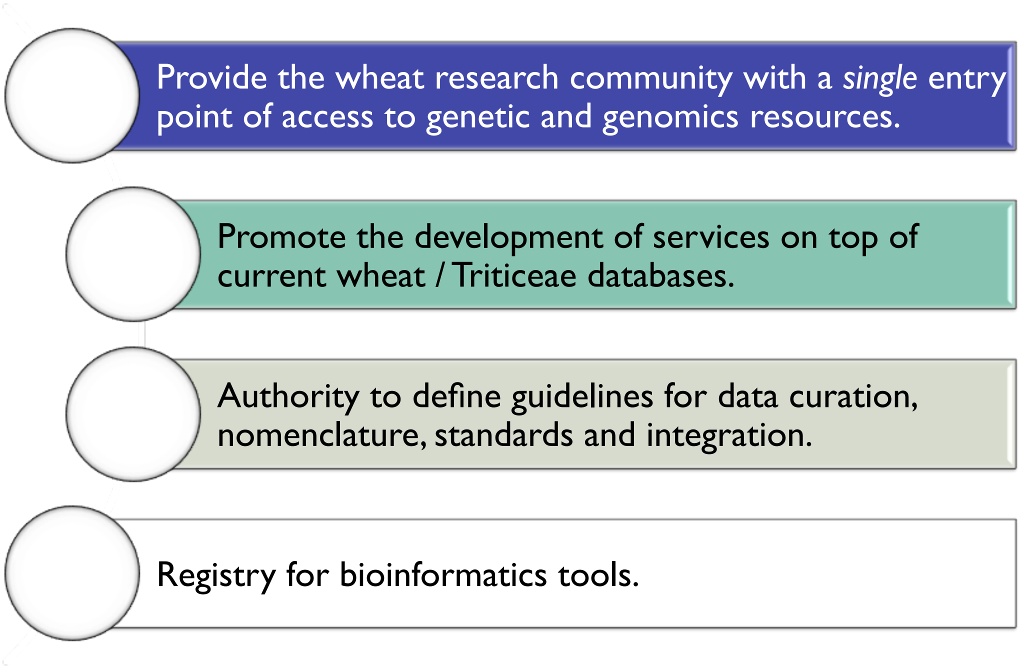
Advance training for current and next generation of scientists

TGAC
The Genome Analysis Centre™

Vicky Schneider

Wheat Information System

Expert Working Group within the Wheat Initiative.



Provide the wheat research community with a *single* entry point of access to genetic and genomics resources.

Promote the development of services on top of current wheat / Triticeae databases.

Authority to define guidelines for data curation, nomenclature, standards and integration.

Registry for bioinformatics tools.

UK Agri-Tech Strategy

- Agri-Informatics Centre for Sustainability Metrics
- Consortium led by Farming Futures (Colin Merritt) group
  - ➡ NIAB, EMR, IBERS, EBI, The Cool Farm Alliance, Fera
  - ➡ CAISM - Centre for Agriculture Informatics and Sustainability Metrics

Stuart Cathpole