
# The Genome Analysis Centre

*Building Excellence in Genomics and Computational Bioscience*

TGAC

The Genome Analysis Centre

# *Wheat genome sequencing: an update from TGAC*


***Sequencing Technology Development***
***now***
***Plant & Microbial Genomics***
***Group Leader***

***Matthew Clark***
***matt.clark@tgac.ac.uk***



BBSRC bioscience for the future

Greater Norwich Development Partnership

South Norfolk COUNCIL

NORWICH City Council

Norfolk County Council

# Project themes

1. ***Genome sequencing***

   **Improve genome**

2. ***Sequence 7 arms***

   **Minimal Tiling Paths**

3. ***Undercover new genetic variation***

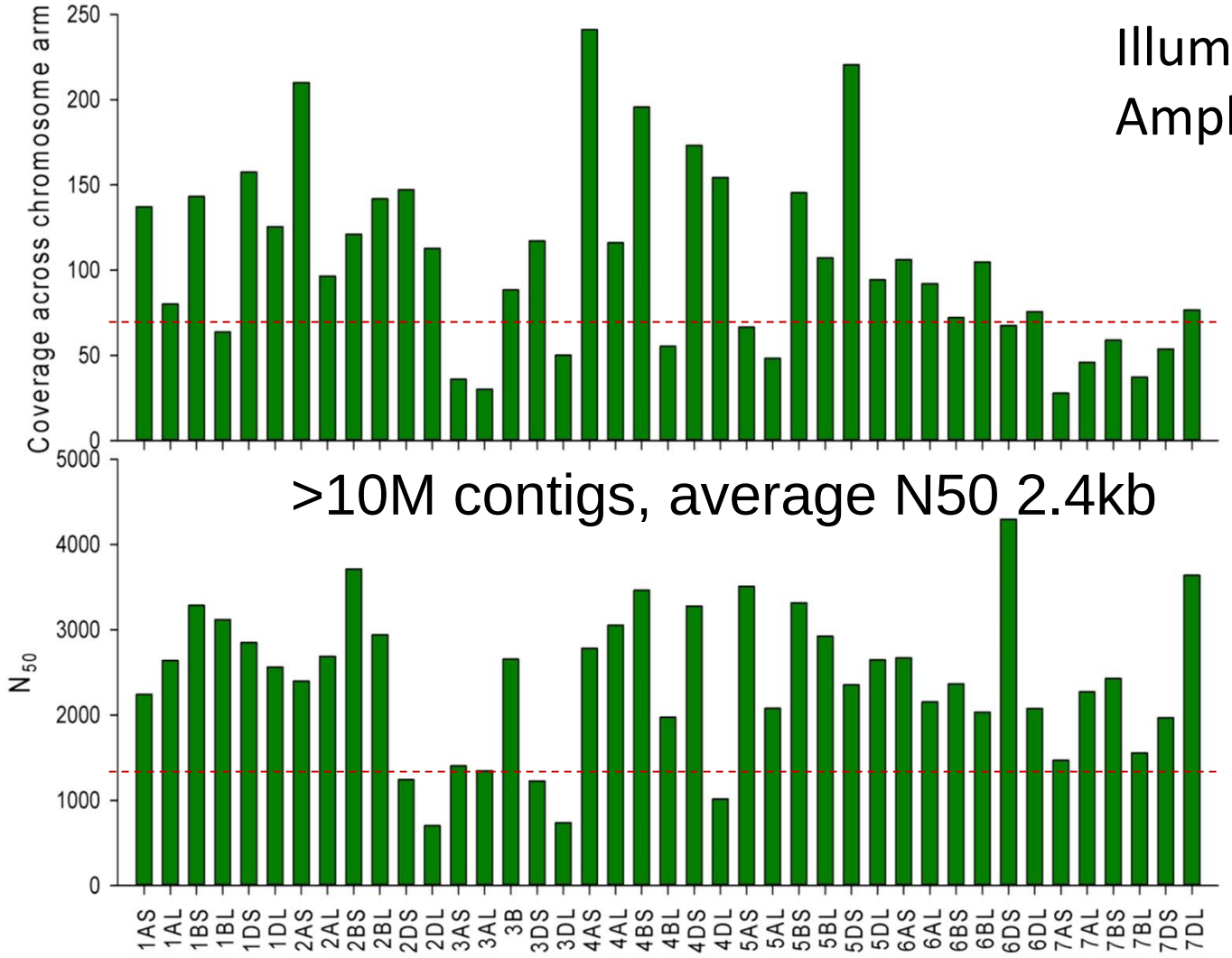
   **TILLING**

**Freely available**

**Easy to browse, search & download etc.**

# Chromosome arm assemblies

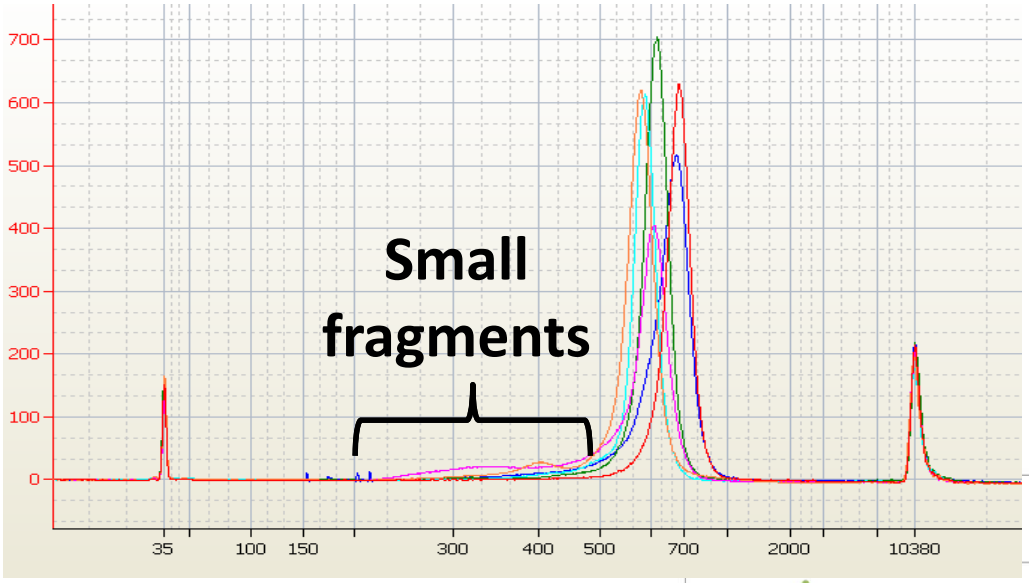
Illumina sequence,
Amplified DNA




>10M contigs, average N50 2.4kb

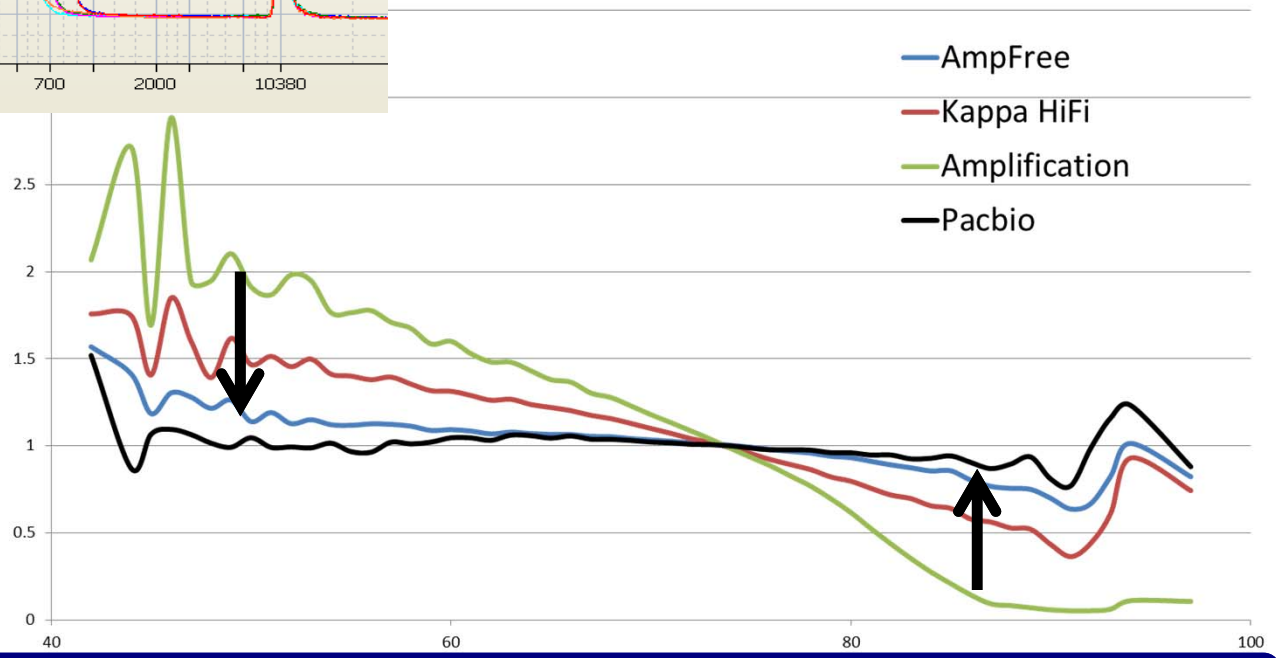
International
Wheat Genome
Sequencing
Consortium

# Improving the genome

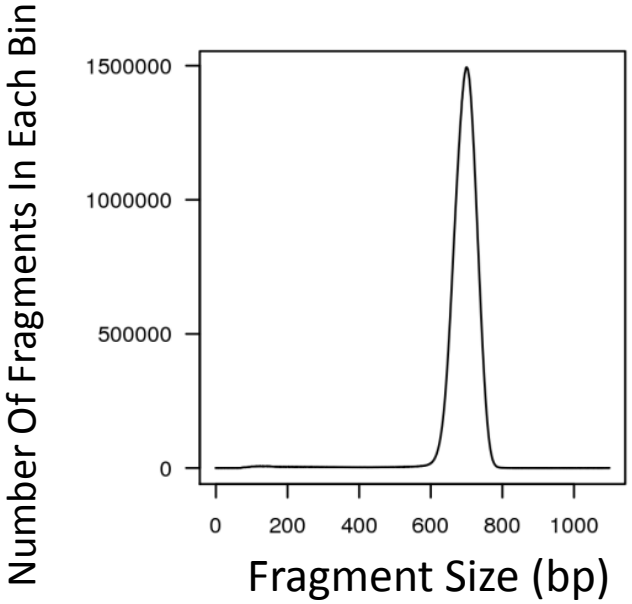


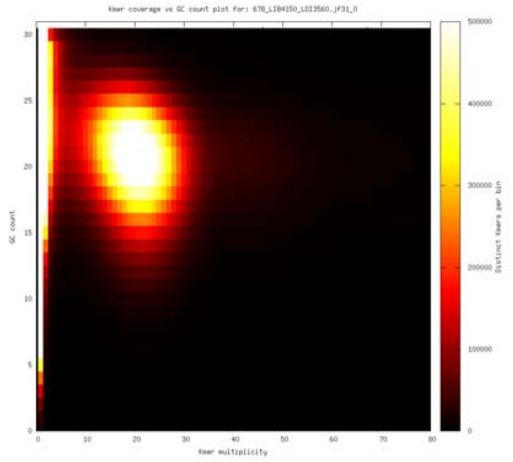

**Wrong size> bad assembly**

**GC bias >**
**Bad assembly**



# Improving the genome

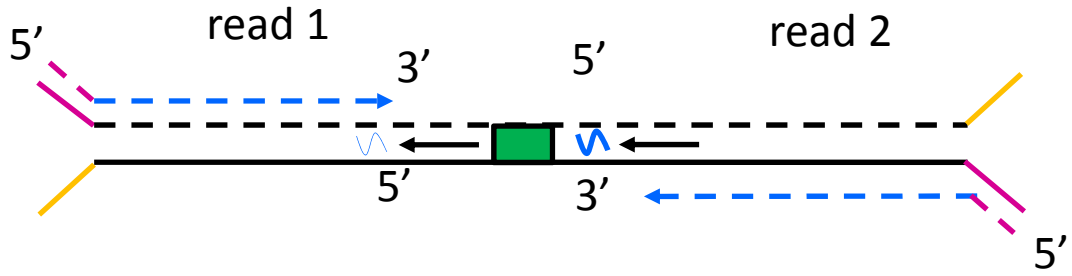


| Read Type | Machine | Read Length | Fragment Size (bp) | Number Of Bases (Gbps) | Read Coverage |
|---|---|---|---|---|---|
| pair end | MiSeq | 2 x 250 bp | 700 | 12.9 | 0.76x |
| pair end | HiSeq 2500 | 2 x 150 bp | 700 | 131.7 | 7.74x |
| pair end | HiSeq 2000 | 2 x 100 bp | 700 | 335.3 | 19.73x |
| mate pair | MiSeq | 2 x 250 bp | 6500 | 12.7 | 0.75x (9.67x*) |

*physical span coverage

# Improving the genome



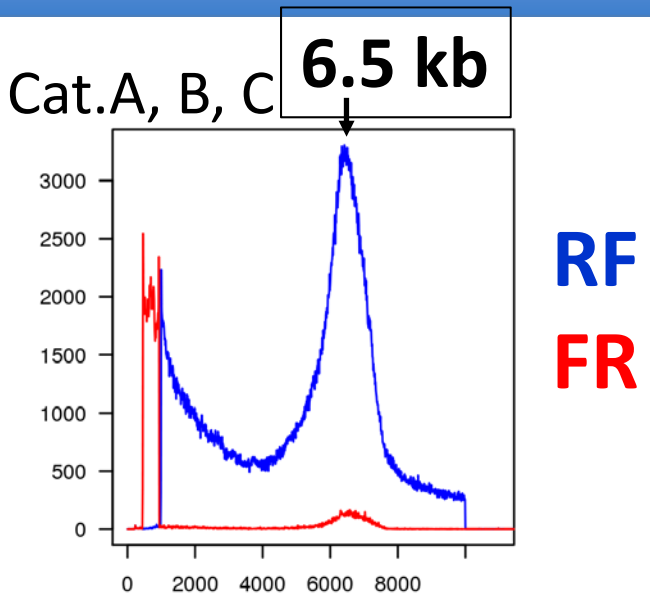

## Nextclip tool (submitted) to classify Nextera mate pairs

| | |
|---|---|
| Adaptor in both reads (A) | 12.64 % |
| Adaptor in read2 only (B) | 24.66 % |
| Adaptor in read1 only (C) | 25.76 % |
| Adaptor not in any read (D) | 18.18 % - mate pairs? |
| Relaxed Category A (E) | 0.17 % |

**63% Mate pairs**

Number Of Fragments In Each Bin

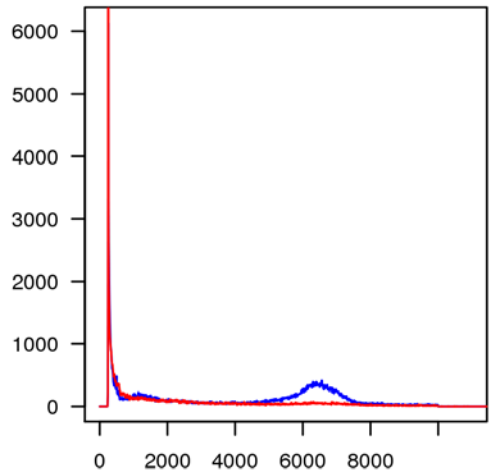
Cat.A, B, C




RF

FR

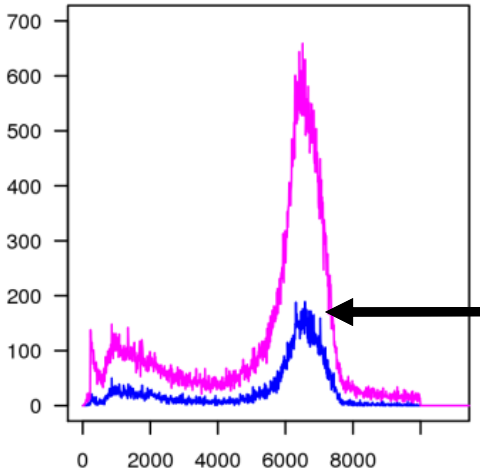
Cat.D




RF

FR

Cat.A, B, C - unique reads



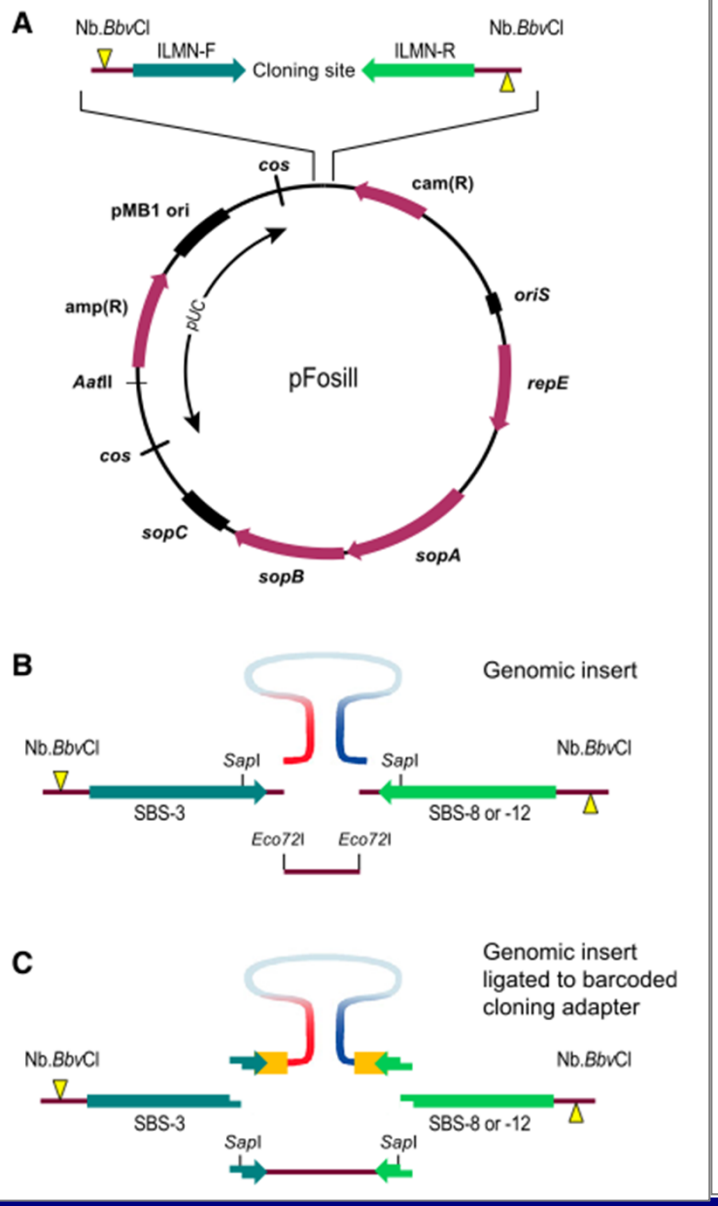

**RF both unique**

**RF one unique**

3-5% of reads map uniquely @ 2*250bp

# 40kb fosmid mate pairs – Broad "Fosills"

TGAC The Genome Analysis Centre™



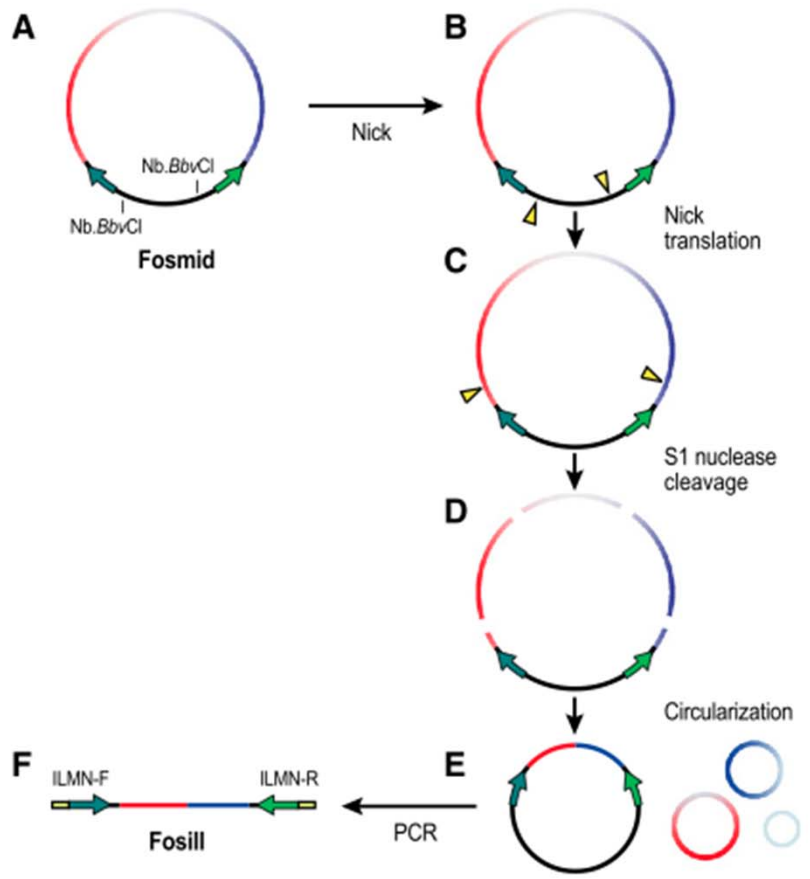
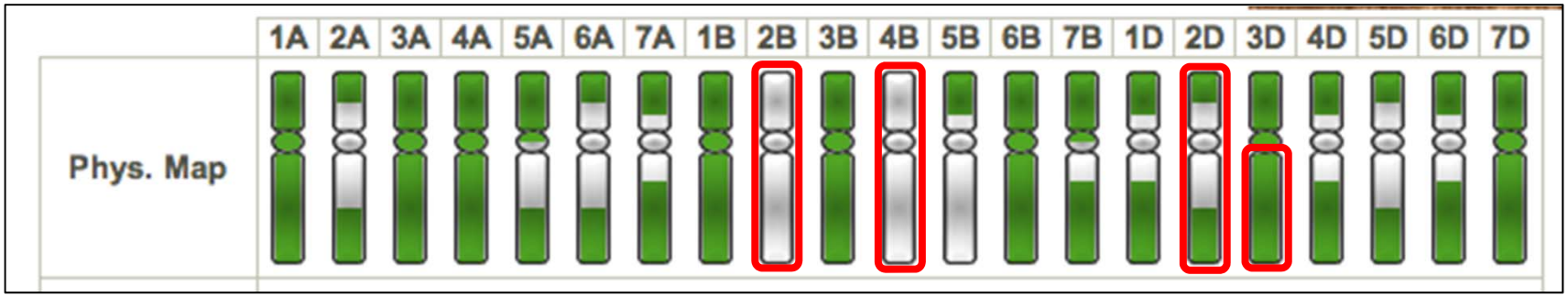




**Figure 2.** Conversion of a Fosmid library to an Illumina-compatible *Fosill* jumping library. (*A,B*) The two Nb.BbvCI sites in the vector are nicked. (C) The nicks are translated in opposite directions into the cloned insert. (*D*) The insert is cleaved at the two translated nicks as well as at nicks originating at any BbvCI sites within the genomic DNA sequence. (*E*) Fragments are circularized by intramolecular ligation. (*F*) Recircularized vector molecules serve as templates for inverse PCR with full-length Illumina enrichment primers that include the sequences required for bridge-amplification and paired-end sequencing of the coligated termini of the original Fosmid insert on the Illumina flow cell.

**But 10x coverage needs 100 ligations**

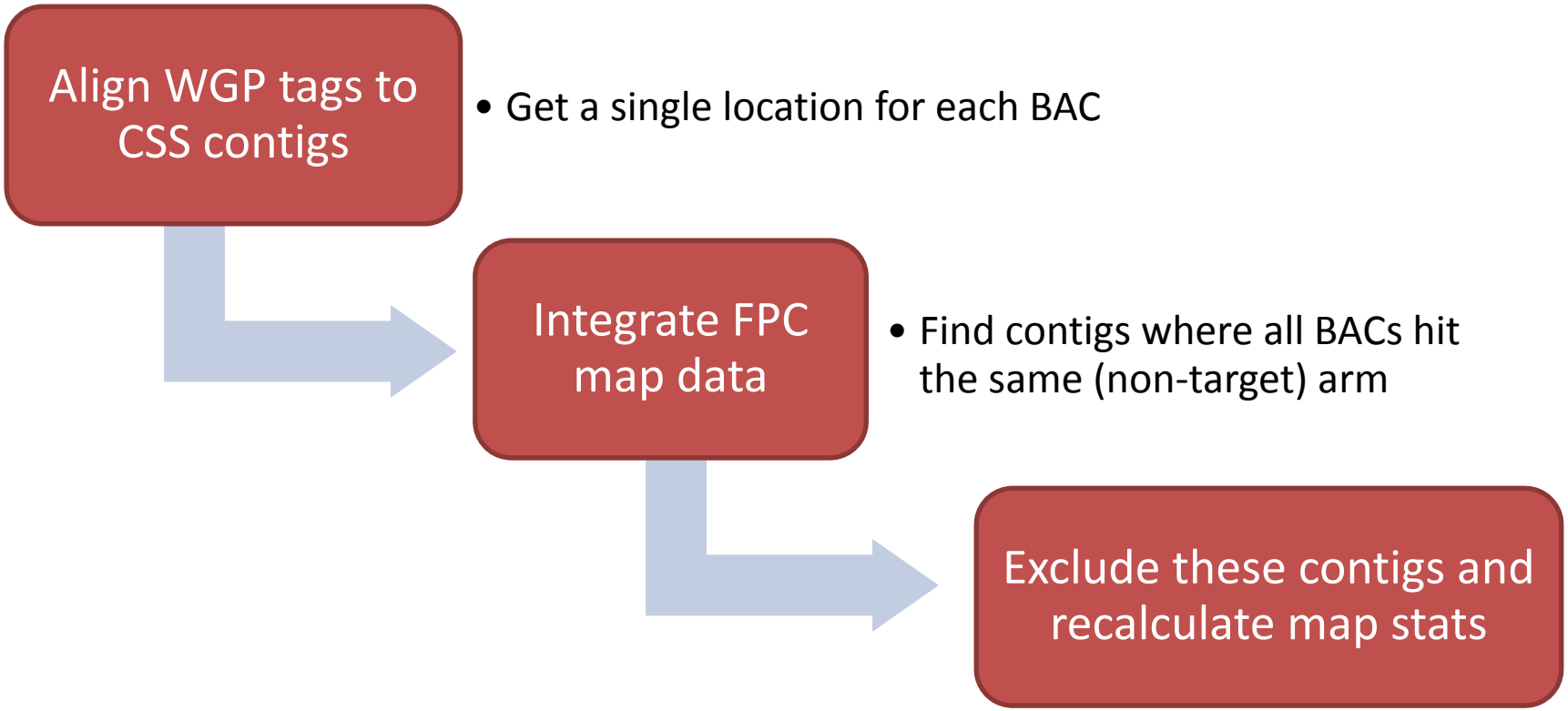
**Mike Bevan JIC**

# 2. Sequencing 7 Chromosome arms

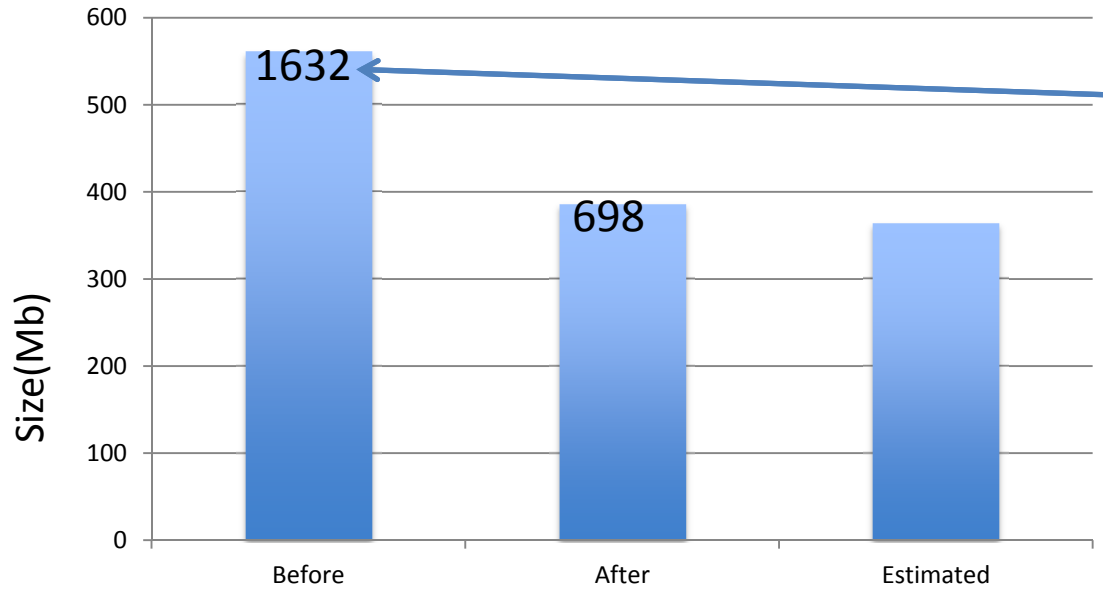



1. 3DL BACs > LTC MTP> BAC DNAs made.
2. 2DS & 2DL BAC libraries profiled, problems.
3. 2BS, 2BL and 4BL BAC constructed.
4. 4BS remains to be constructed.

# 2DS & 2DL contamination contigs

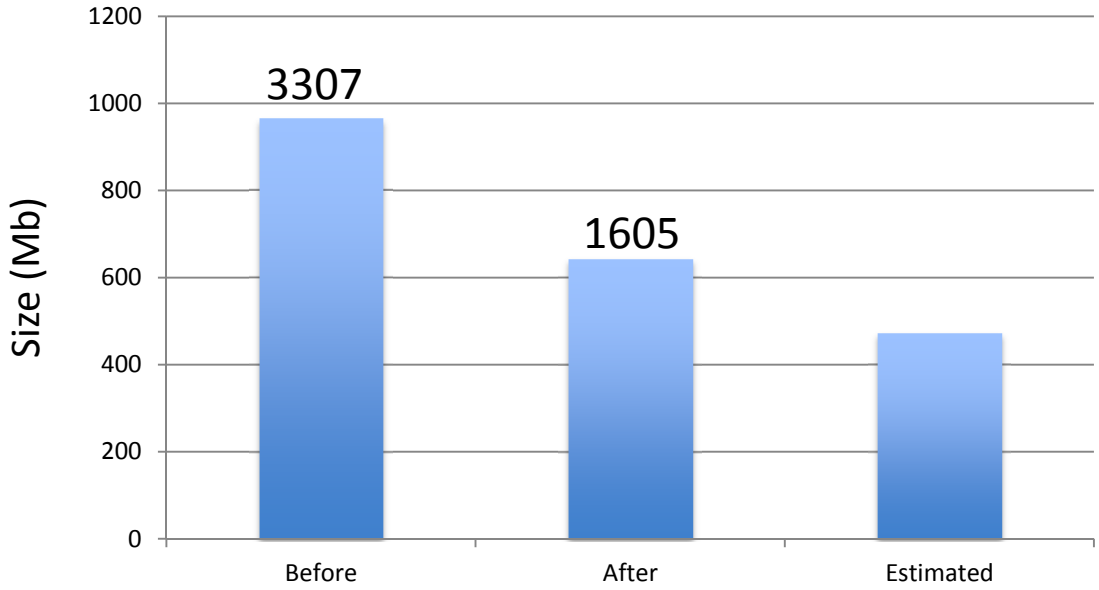


## 2DS

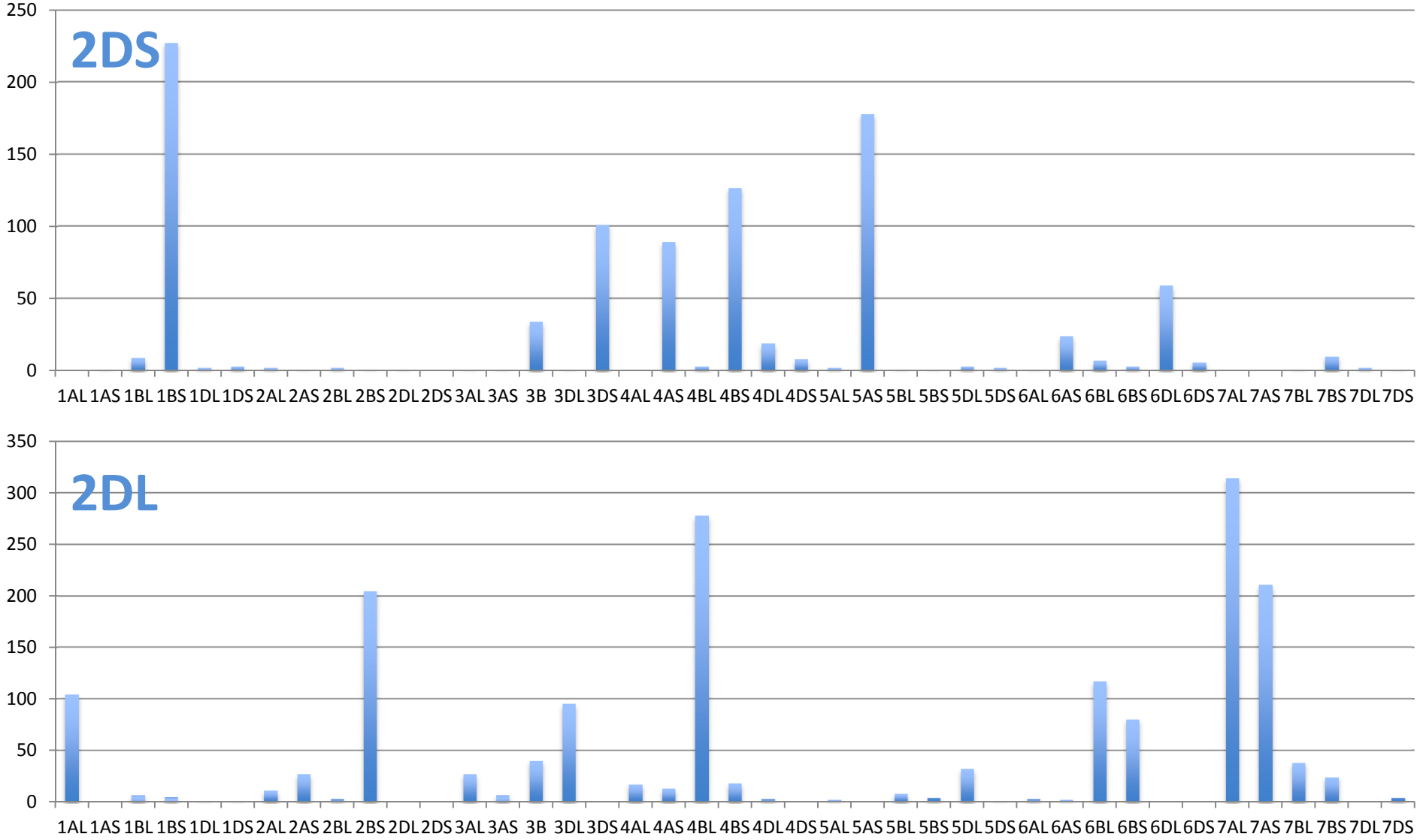



Number of contigs

## 2DL



# The origin of contamination



# How to sequence 10,000 BACs?

1. ***NGS-ready BAC DNA***

   **High throughput & in microtitre plates**

2. ***Cheap indexed libraries***

   **<<< $100/library**

3. ***High throughput***

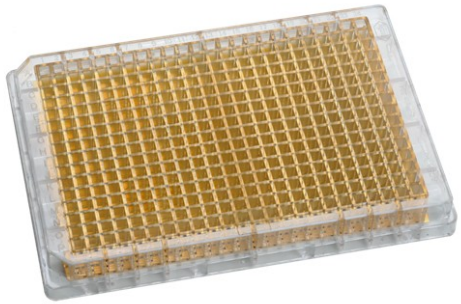
   **Process 10k BACs (chromosome) in weeks**

4. ***Good assemblies***

   **Few contigs per BAC, ideally one.**

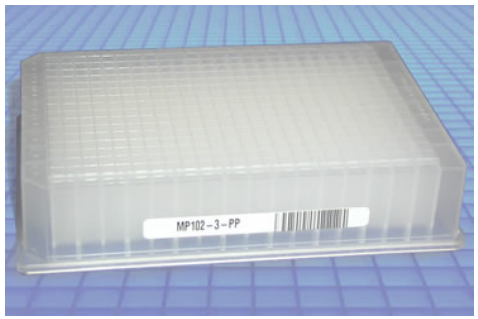
# NGS ready BAC DNA

**Library plate**

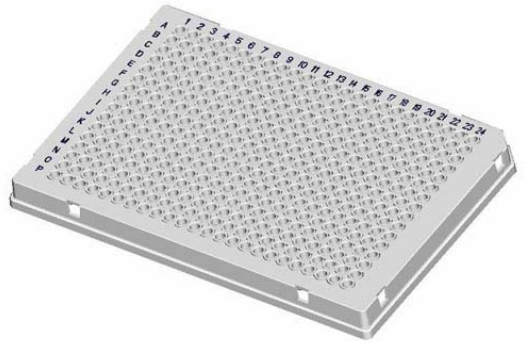


*Copy and grow*

**Deep culture plate**




*Bead based DNA prep*

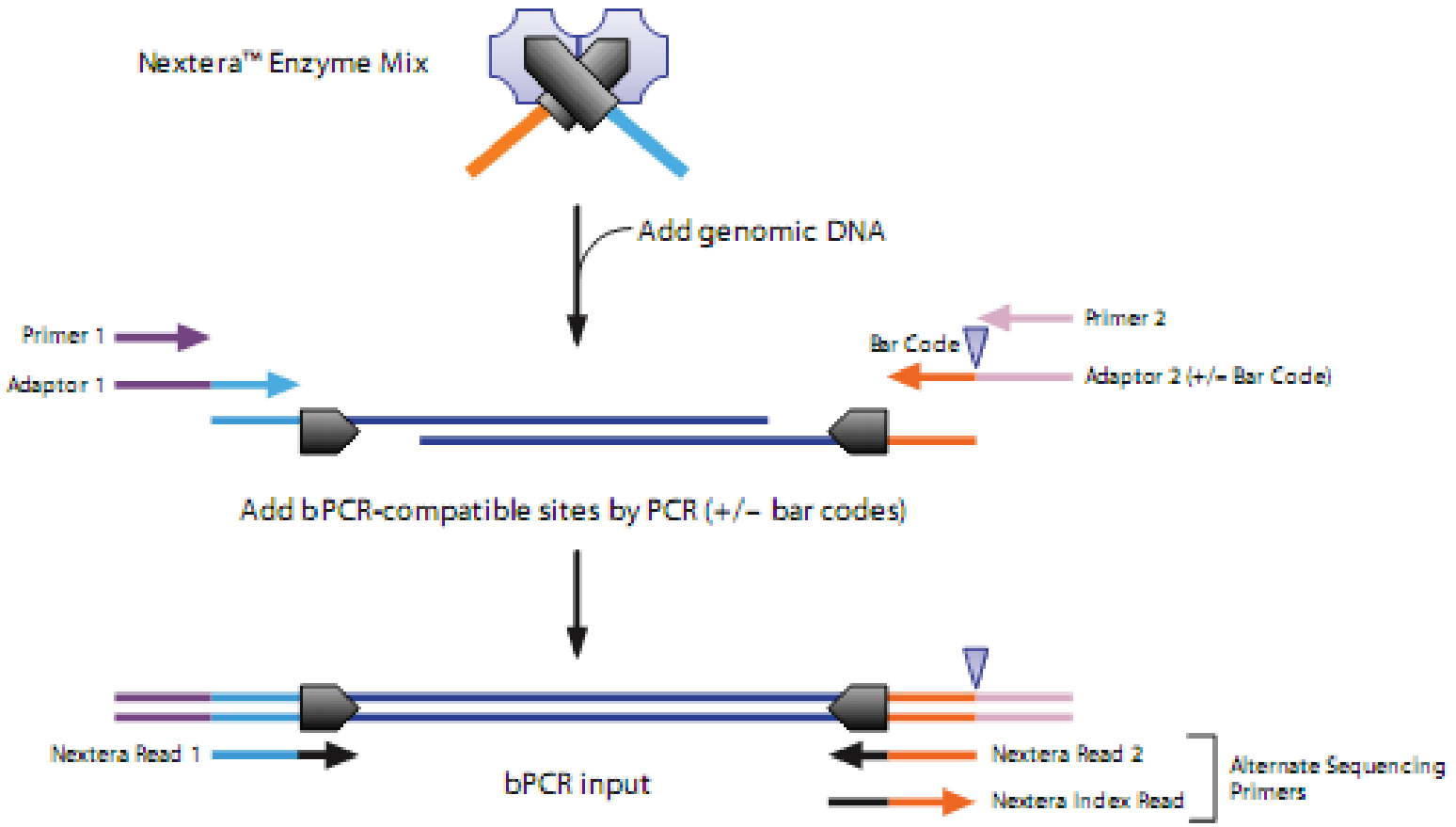


**~100ng BAC DNA**

*Removal gDNA*

**95+% pure BAC DNA**

# Cheap indexed libraries



# Cheap indexed libraries

**95% pure BAC DNA**

"Tagmentation"

Add indexed primers & PCR

Size selection

$7 (FEC)/BAC = DNA & library

**$8 (FEC)/BAC = sequencing @384-plex**

# Cheap indexed libraries

Illumina Sequencing Analysis Viewer 1.8.4 - 120412_M00392_0009_AMS0006782-00300

Sequencing Analysis Viewer

Run Folder: D:\illumina\MiSeqTemp\120412_M00392_0009_AMS0006782-00300 Browse Refresh

Analysis | Imaging | Summary | Tile Status | TruSeq Controls | Indexing

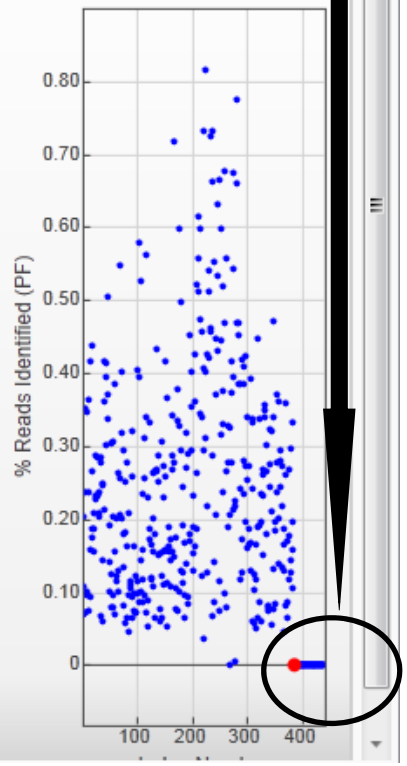
Lane 1

0.0003% incorrect

**Reads mapped to Index Id**

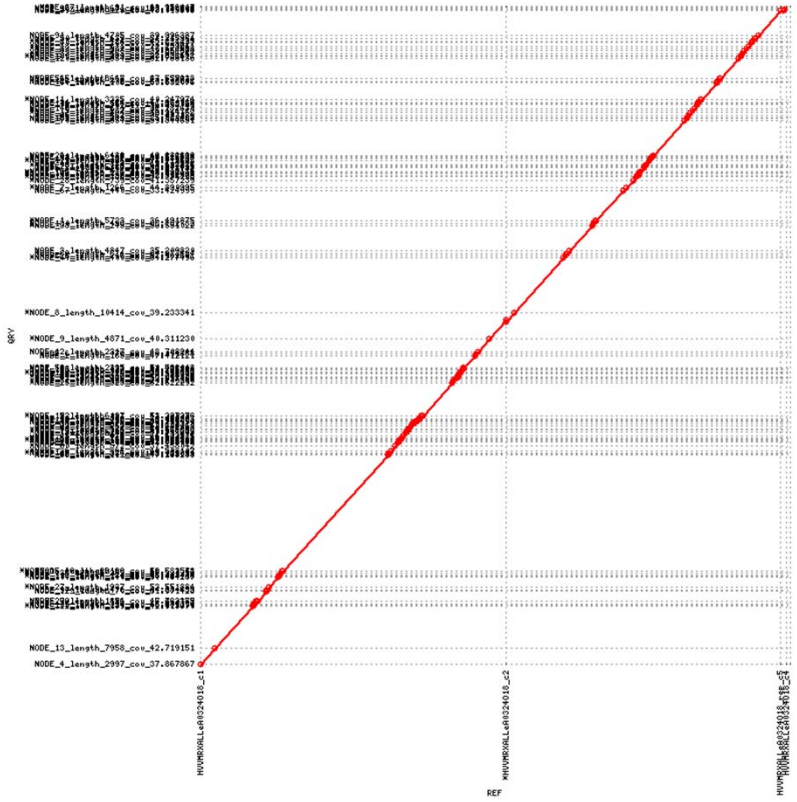
| Total Reads | PF Reads | % Reads Identified (PF) | CV | Min | Max |
|---|---|---|---|---|---|
| 6143396 | 4644384 | 96.4731 | 0.7626 | 0 | 0.8172 |

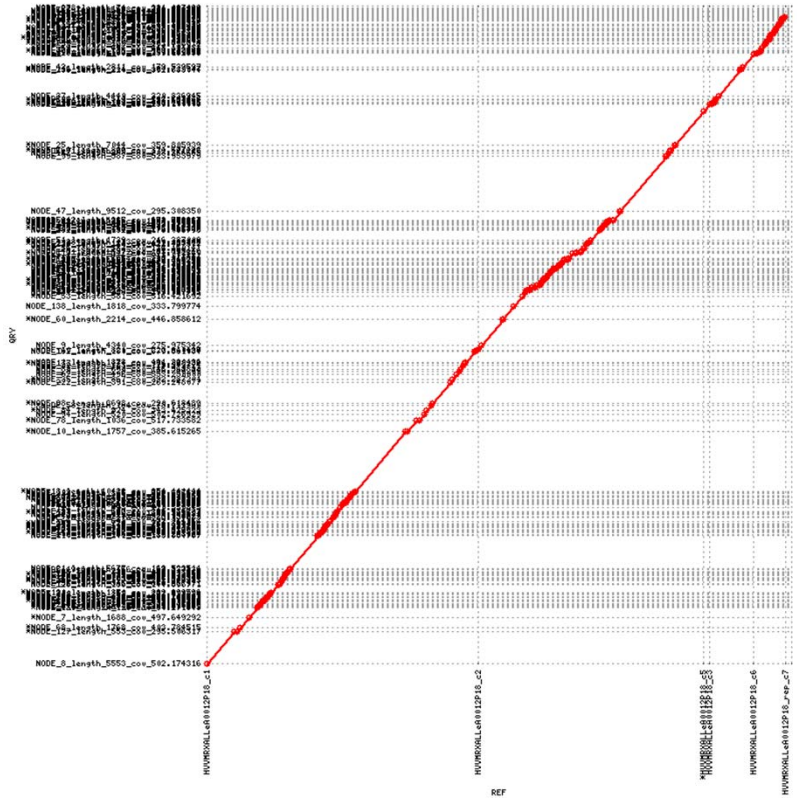
| Index Number | Sample Id | Project | Index 1 (I7) | Index 2 (I5) | % Reads Identified (PF) |
|---|---|---|---|---|---|
| 375 | NextAdapt2_IDX_L9D4_375 | Barley | GTTGCATCG | | 0.1186 |
| 376 | NextAdapt2_IDX_L9D4_376 | Barley | TAACTCTAC | | 0.1787 |
| 377 | NextAdapt2_IDX_L9D4_377 | Barley | TAAGTATTG | | 0.2681 |
| 378 | NextAdapt2_IDX_L9D4_378 | Barley | TAATGGAAG | | 0.1276 |
| 379 | NextAdapt2_IDX_L9D4_379 | Barley | TACCGTACC | | 0.2982 |
| 380 | NextAdapt2_IDX_L9D4_380 | Barley | TAGGCGGCG | | 0.1437 |
| 381 | NextAdapt2_IDX_L9D4_381 | Barley | TAGGTTAGG | | 0.1958 |
| 382 | NextAdapt2_IDX_L9D4_382 | Barley | TAGTCCTGG | | 0.3341 |
| 383 | NextAdapt2_IDX_L9D4_383 | Barley | TAGTTATAT | | 0.1068 |
| 384 | NextAdapt2_IDX_L9D4_384 | Barley | TATCAAGCC | | 0.1589 |
| 385 | NextAdapt2_IDX_L9D4_385 | Barley | TATCATTAT | | 0 |
| 386 | NextAdapt2_IDX_L9D4_386 | Barley | TATGCTCGC | | 0 |
| 387 | NextAdapt2_IDX_L9D4_387 | Barley | TATGGATAA | | 0 |
| 388 | NextAdapt2_IDX_L9D4_388 | Barley | TATTAGAGT | | 0 |
| 389 | NextAdapt2_IDX_L9D4_389 | Barley | TCAACTTAC | | 0 |
| 390 | NextAdapt2_IDX_L9D4_390 | Barley | TCAGCGATT | | 0 |
| 391 | NextAdapt2_IDX_L9D4_391 | Barley | TCATGCGCC | | 0 |
| 392 | NextAdapt2_IDX_L9D4_392 | Barley | TCCGCCGAT | | 0.0001 |
| 393 | NextAdapt2_IDX_L9D4_393 | Barley | TCCTCGAAC | | 0 |
| 394 | NextAdapt2_IDX_L9D4_394 | Barley | TCCTCTCGG | | 0 |
| 395 | NextAdapt2_IDX_L9D4_395 | Barley | TCCTTCTTG | | 0.0003 |
| 396 | NextAdapt2_IDX_L9D4_396 | Barley | TCGCCTGCC | | 0.0001 |



TGAC
The Genome Analysis Centre™

# Cheap indexed libraries





IPK- 454 (x) versus TGAC-Illumina (y)

TGAC assemblies Velvet N50 = ~5kb

# High throughput

A. Scale up – 1 plate to many

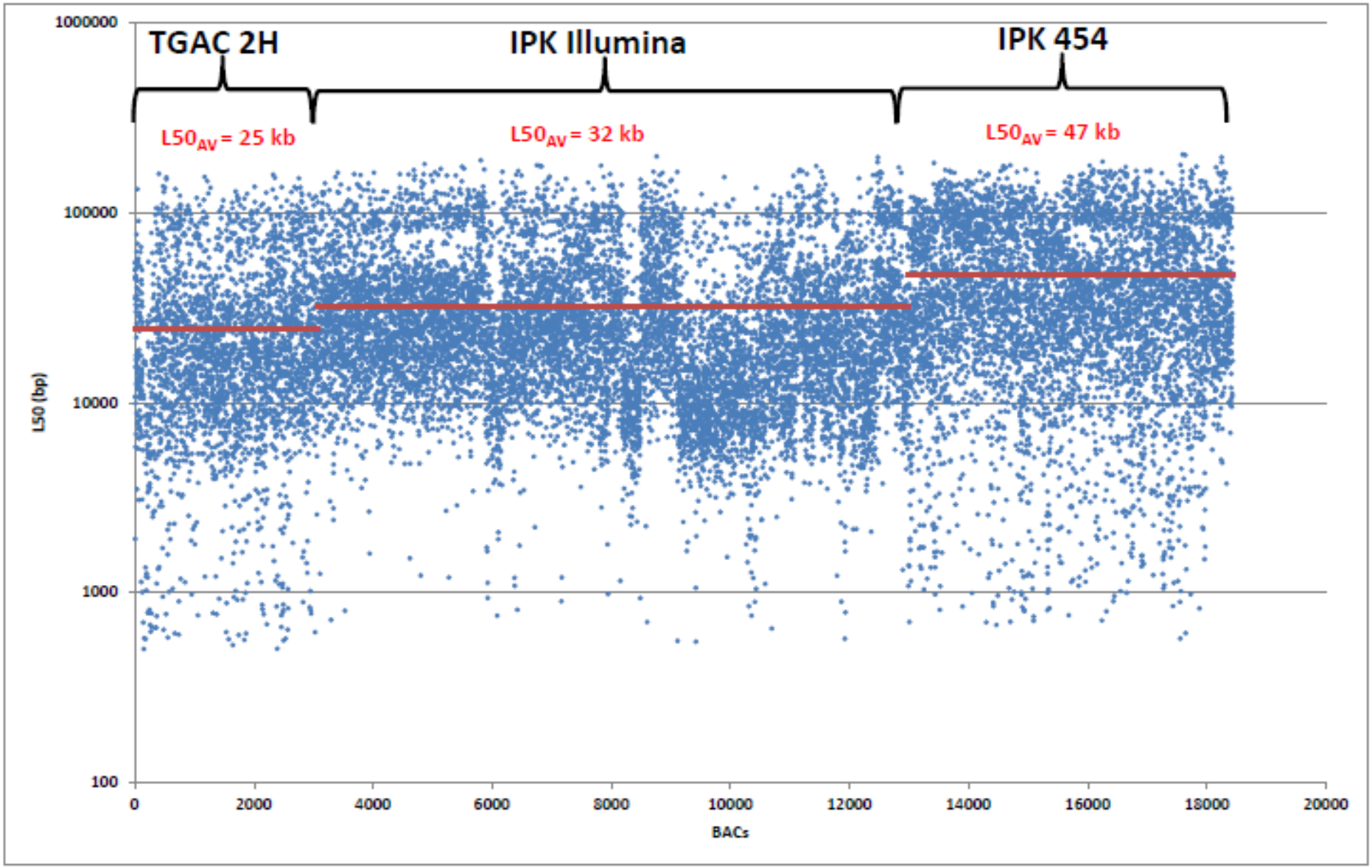
B. Fill an Illumina HiSeq 2000

C. 16 lanes @ 384plex = 6,144 BACs

D. DNA and libraries + 11 days to sequence

E. 6,144 2H MTP BACs in 4 weeks, from bacteria to sequence

# IPK assembly using CLC

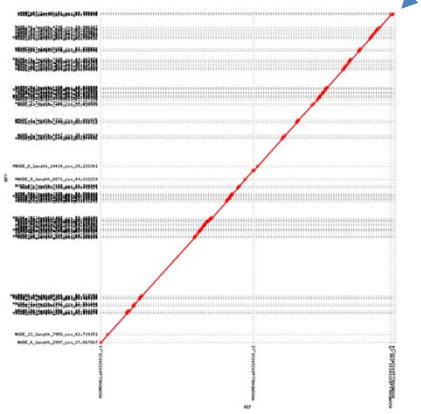


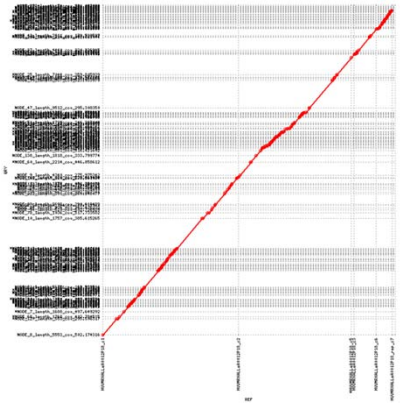
# Improved assemblies

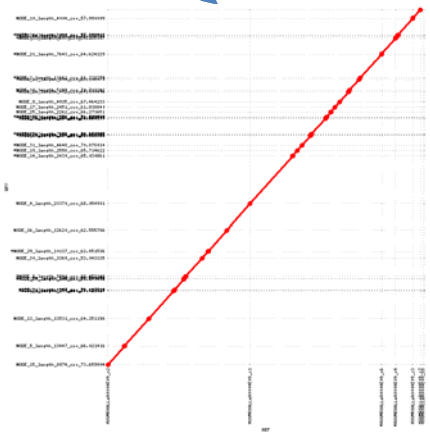
**384 BACs**

Nextera mate pair (384 BAC pool)

Assign to BACs
(map to PE assemblies)



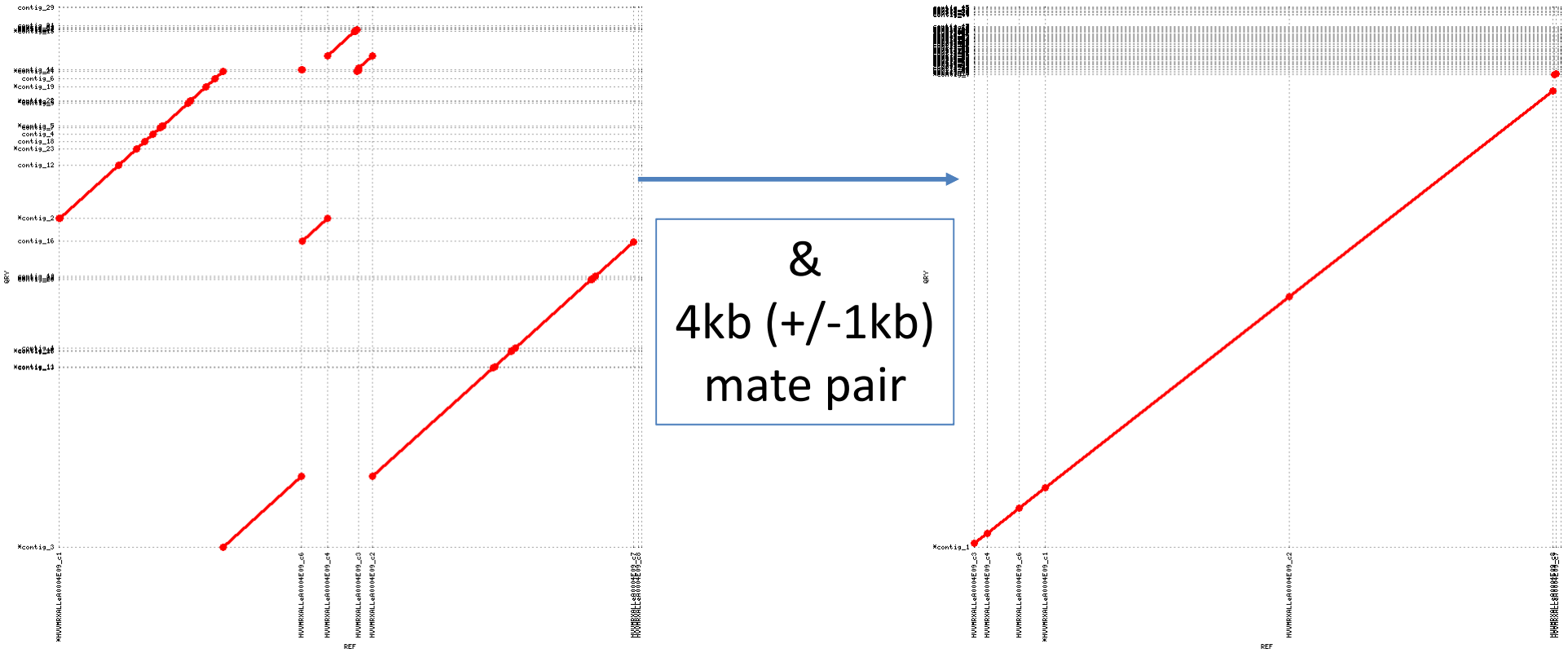




# Improved assemblies using LMPs

TGAC

## Improved assemblies




## CLC assemblies

Deepali Vasoya. The Genome Analysis Centre

# 3. Undercover new genetic variation

1. Population of (EMS) mutagenised plants;
2. High throughput screen to identify mutations in a gene of interest.
3. Two populations, Kronos (durum) and Cadenza (bread)
4. Exome capture
5. Low-cost sequencing

Reverse Genetics

# Wheat capture for TILLing

TGAC

*T. urartu* and *T. turgidum* RNAseq *de novo* assembled (CLC)

↓

*+ T. aestivum*, wheatified barley models and popular wheat genes

↓

CDS models aligned to CSS arms (exonerate) - filtered by length and %id

↓

Exons padded with +/-30bp genomic sequence

UCDAVIS UNIVERSITY OF CALIFORNIA

John Innes Centre

BBSRC bioscience for the future

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

# Wheat capture stats

TGAC

**Capture contains 82,091 transcripts, in 286,799 exons, totalling 84 Mb**

**200,584 exons were padded +/-30bp (70%)**

**11,344 transcripts were included without a genomic model**

**Alpha design currently being synthesised**

# Exome capture of 1,536 4x and 6x mutants

Submitted >250,000 exons (82+k transcripts) for Nimblegen design

Sequencing of first mutants by end of the year

**By end of 2014*:***

- *in silico* access to TILLING alleles in 4x and 6x wheat
- >85% probability of a knock-out allele

***Access to mutants***

- We plan to hold mirror collection of seeds at UC Davis and JIC Germplasm Collection
- Mutants will be free from any IP for the mutations people find
- We plan to charge a small fee for ~10 seeds of each mutant to maintain collections

UC DAVIS
UNIVERSITY OF CALIFORNIA

John Innes Centre

ROTHAMSTED RESEARCH

# Acknowledgements

TGAC

## Computational Genomics

- Paul Bailey
- Jon Wright
- Bernardo Clavijo
- Dharanya Sampath
- Sarah Ayling

## Plant & Microbial Genomics

- Darren Heavens
- Deepali Vasoya

## Library & Sequencing teams

## Mario Caccamo

International Wheat Genome Sequencing Consortium

John Innes Centre

UC DAVIS UNIVERSITY OF CALIFORNIA

EMBL-EBI

IPK GATERSLEBEN

The James Hutton Institute

BBSRC bioscience for the future

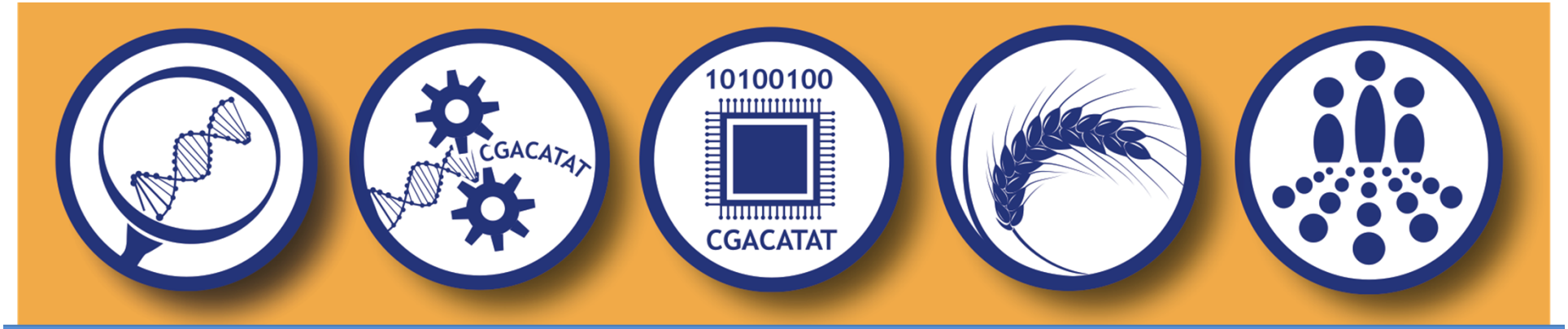
ROTHAMSTED RESEARCH

NORWICH City Council

Greater Norwich Development Partnership

Norfolk County Council

South Norfolk COUNCIL




# Fellowship Programme in Computational Biology

- Bioinformatics and Computational Biology.
- Competitive salary and research support grant.
- Fast application process, call open until posts are filled.
- Up to five years: for early career scientists who want to become scientific leaders in a dynamic research environment!

www.tgac.ac.uk/fellowship

Apply now!!!

TGAC
The Genome Analysis Centre™

BBSRC

Greater Norwich Development Partnership