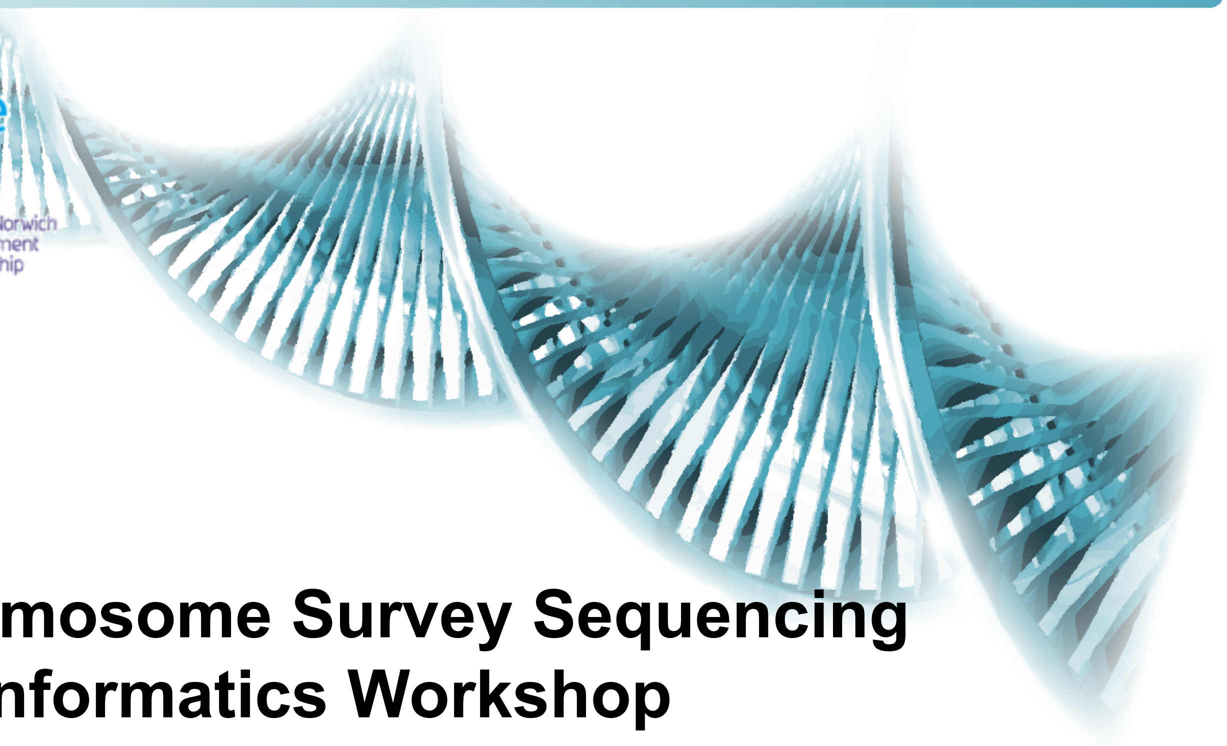
BBSRC Genome Analysis Centre
In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium



Wheat Chromosome Survey Sequencing Bioinformatics Workshop

Mario Caccamo – Jon Wright
Bioinformatics division
The Genome Analysis Centre
wheatdcc.tgac.bbsrc.ac.uk

BBSRC bioscience for the future · South Norfolk COUNCIL · NORWICH City Council · Norfolk County Council · EEDA East of England Development Agency · Greater Norwich Development Partnership

Bioinformatics Pipelines

BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium

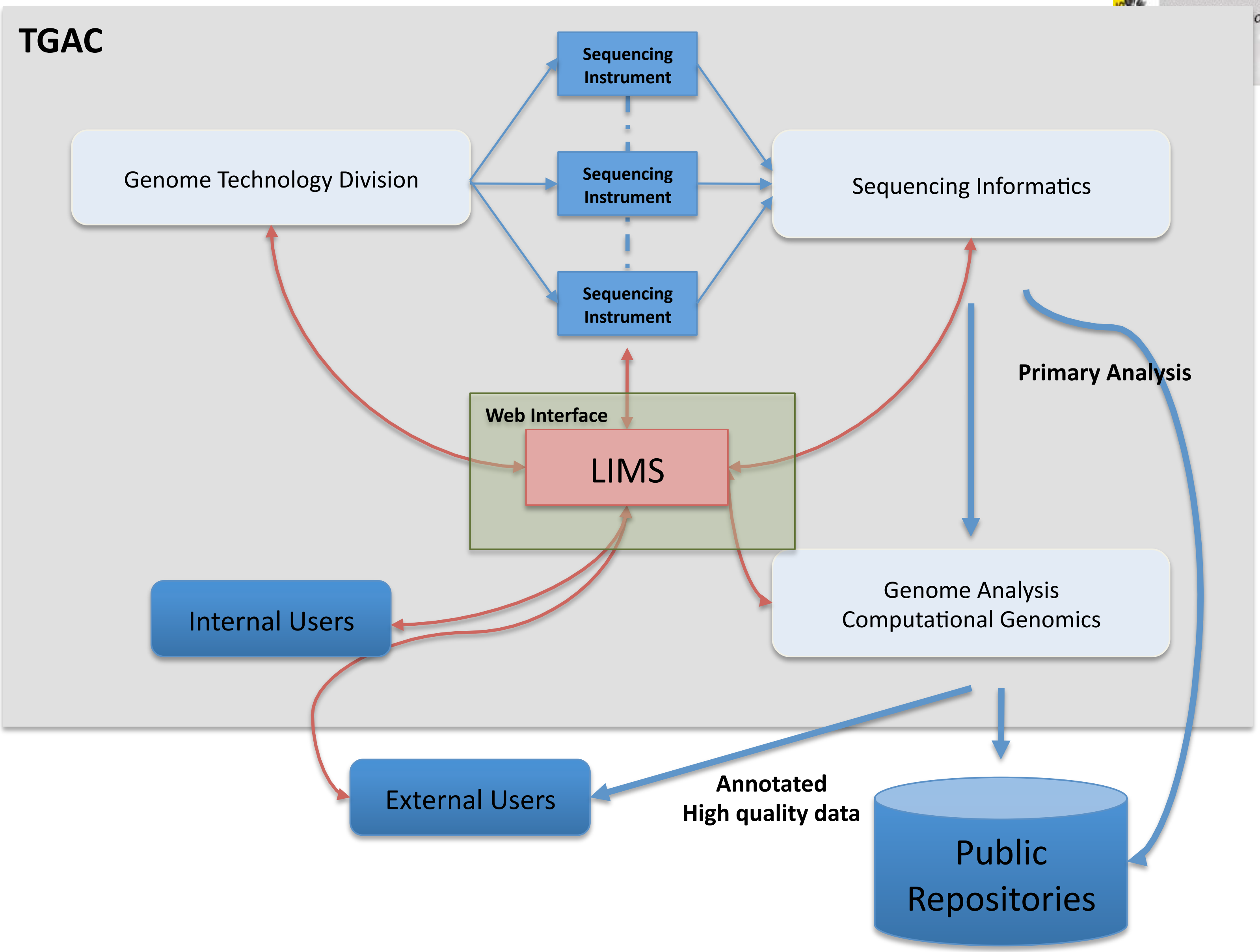
- ***De novo* genome sequencing** and associated analysis

- **Re-sequencing** for variation and population analysis

- **Transcriptome analysis**
  - Studying gene expression levels and patterns
  - Regulatory changes
  - Rare variants and associated expression changes
  - Transcription regulation by epigenetic markers

- **Metagenomics/Metatranscriptomics**
  - Analysis of environmental samples to identify new genes and pathways e.g. in soil or the human gut microbiome

Data Analysis Pipeline



TGAC

Genome Technology Division

Sequencing Instrument

Sequencing Instrument

Sequencing Instrument

Sequencing Informatics

Web Interface

LIMS

Primary Analysis

Internal Users

Genome Analysis
Computational Genomics

External Users

Annotated
High quality data

Public
Repositories

TGAC Computing Capacity

Phase 1 (Sep ’08 - Mar’10)

100 TB storage capacity mirrored

Linux cluster with 120 computing nodes, ~400 GB RAM for data processing

Phase 2 (Apr’10 - Mar’11)

New Data Centre in B26 (also houses training lab + computing training facility)

0.6 PB storage capacity mirrored

1000 computing nodes; 4 x 256GB RAM

Big memory machine: SGI Altix UV100 (6TB RAM, 576 CPU cores)

Phase 3

Future options - use of HTC facilities, cloud computing?

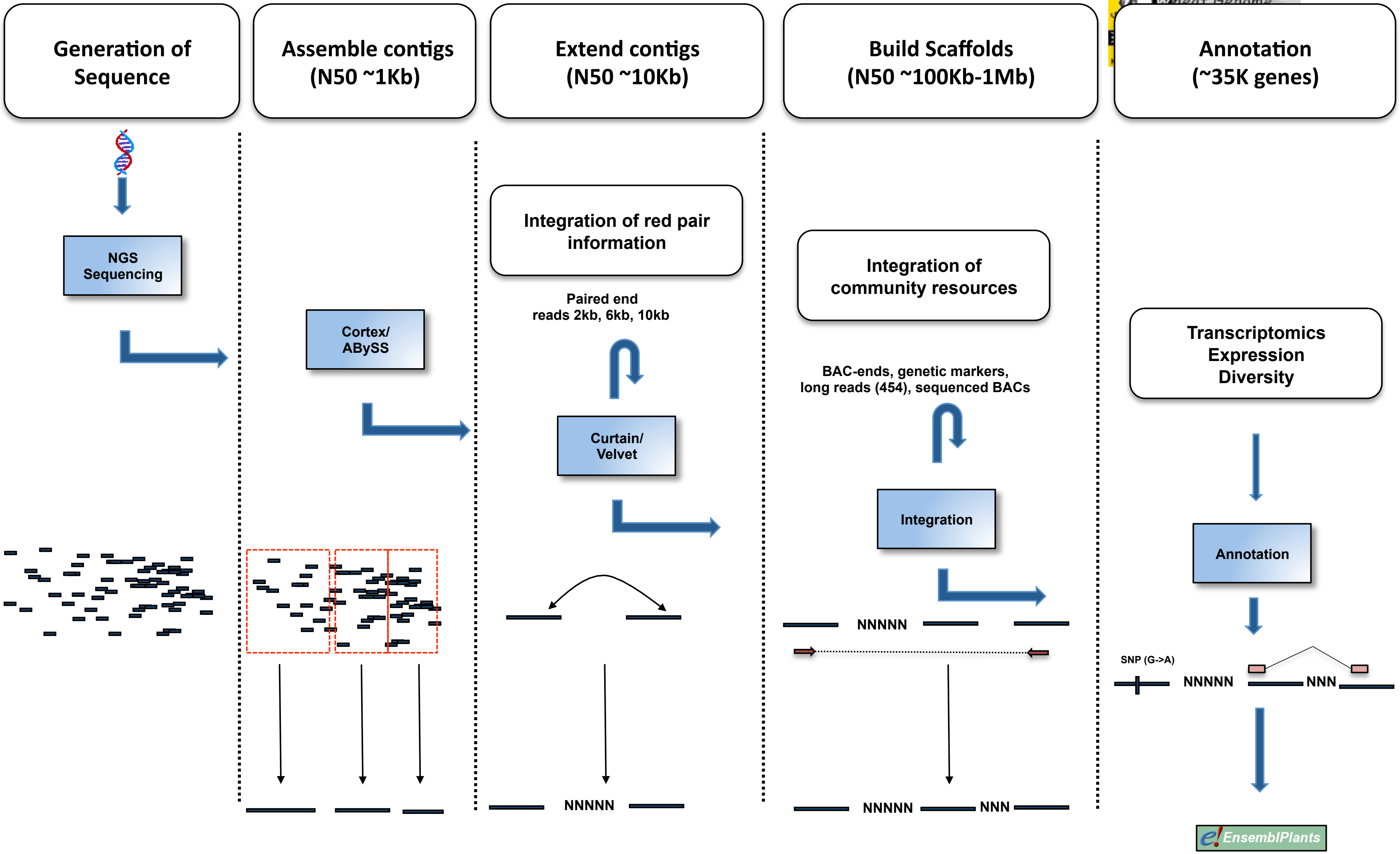
Big Data Challenge

Assembly Pipeline

BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome



Generation of Sequence

Assemble contigs (N50 ~1Kb)

Extend contigs (N50 ~10Kb)

Build Scaffolds (N50 ~100Kb-1Mb)

Annotation (~35K genes)

NGS Sequencing

Cortex/ ABySS

Integration of red pair information

Paired end reads 2kb, 6kb, 10kb

Curtain/ Velvet

Integration of community resources

BAC-ends, genetic markers, long reads (454), sequenced BACs

Integration

Transcriptomics Expression Diversity

Annotation

Agenda

- **Wheat Chromosome Sequencing Survey DCC**
- **Assemblies - theory**
- **Assemblies - practice**

Agenda

- **Wheat Chromosome Sequencing Survey DCC**
- Assemblies - theory
- Assemblies - practice

DCC Role

- Track progress for the submission of the WCSS datasets

- Run general QC checks
  - Base content / dinucleotide
  - Quality scores distribution
  - K-mer frequency
  - Contamination screening

- Run the assemblies
  - ABySS, Cortex, CLC, SGA, others

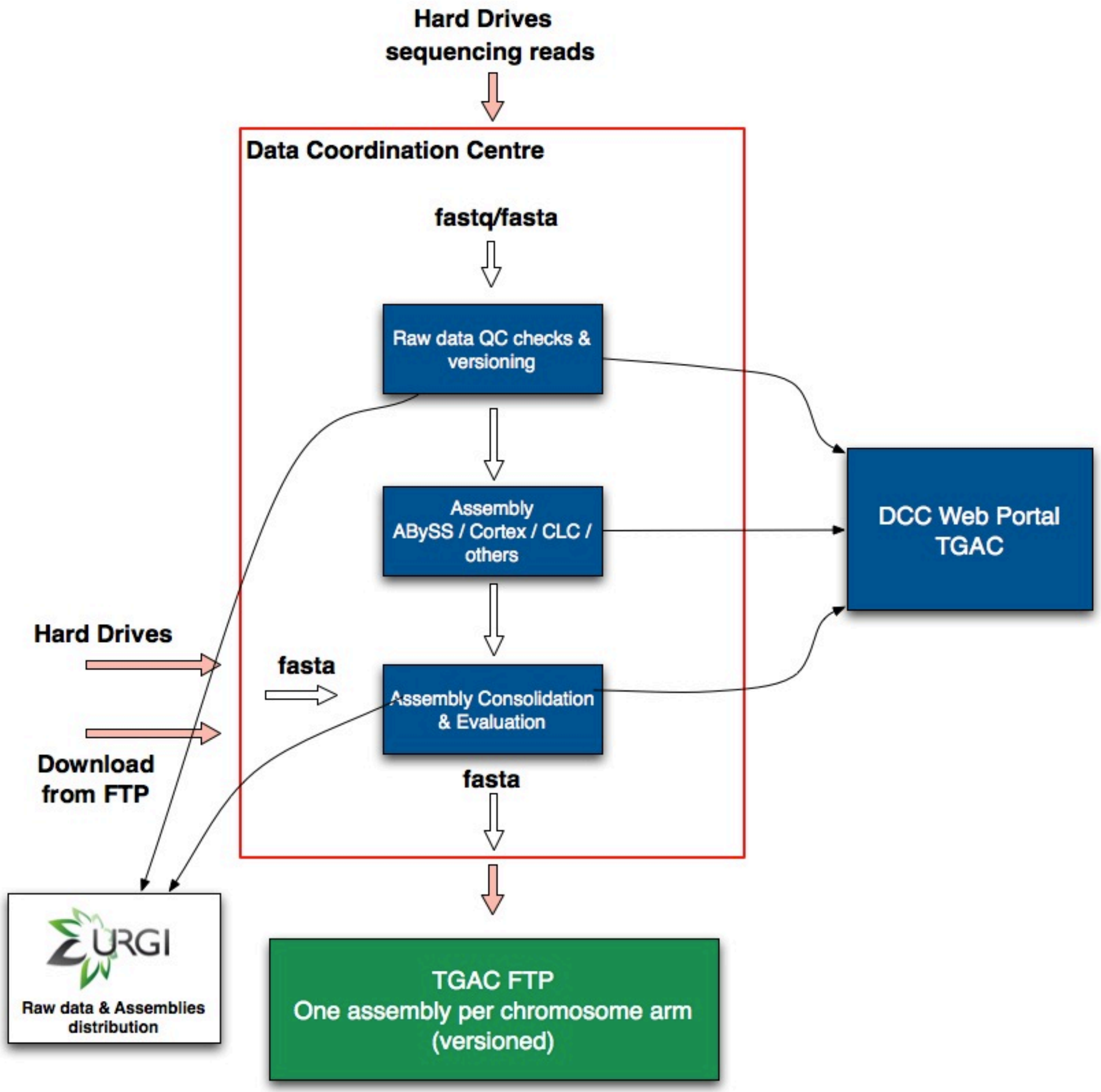
- Consolidate and version the assemblies

- Define the project **data freeze(s)**.

BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium



Hard Drives sequencing reads

Data Coordination Centre

fastq/fasta

Raw data QC checks & versioning

Assembly ABySS / Cortex / CLC / others

Assembly Consolidation & Evaluation

fasta

fasta

DCC Web Portal TGAC

Hard Drives

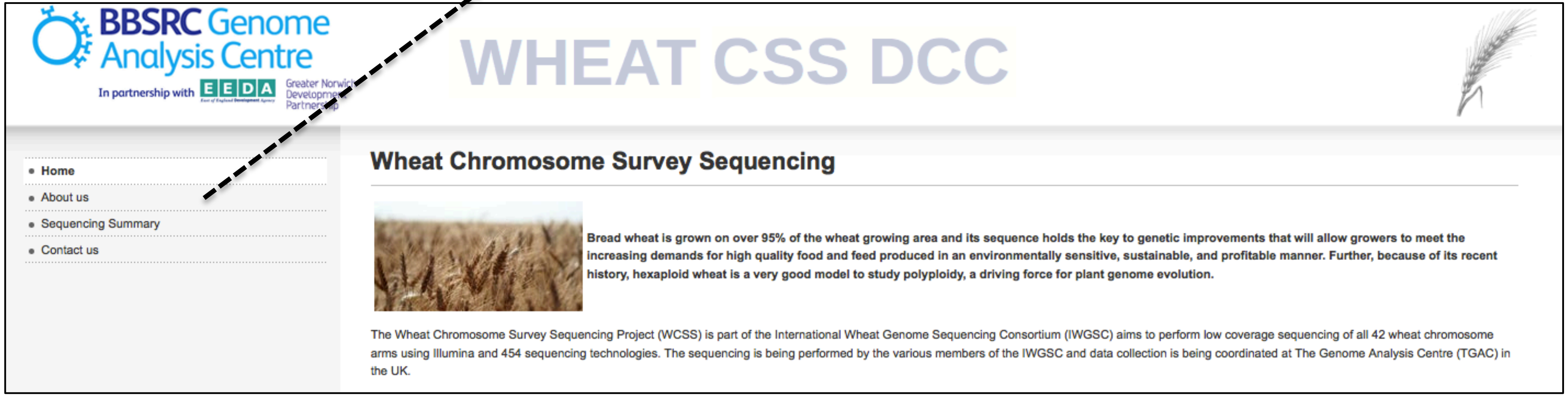
Download from FTP

URGI Raw data & Assemblies distribution

TGAC FTP One assembly per chromosome arm (versioned)

DCC Web Portal

wheatdcc.tgac.bbsrc.ac.uk



WHEAT CSS DCC

Wheat Chromosome Survey Sequencing

Bread wheat is grown on over 95% of the wheat growing area and its sequence holds the key to genetic improvements that will allow growers to meet the increasing demands for high quality food and feed produced in an environmentally sensitive, sustainable, and profitable manner. Further, because of its recent history, hexaploid wheat is a very good model to study polyploidy, a driving force for plant genome evolution.

The Wheat Chromosome Survey Sequencing Project (WCSS) is part of the International Wheat Genome Sequencing Consortium (IWGSC) aims to perform low coverage sequencing of all 42 wheat chromosome arms using Illumina and 454 sequencing technologies. The sequencing is being performed by the various members of the IWGSC and data collection is being coordinated at The Genome Analysis Centre (TGAC) in the UK.

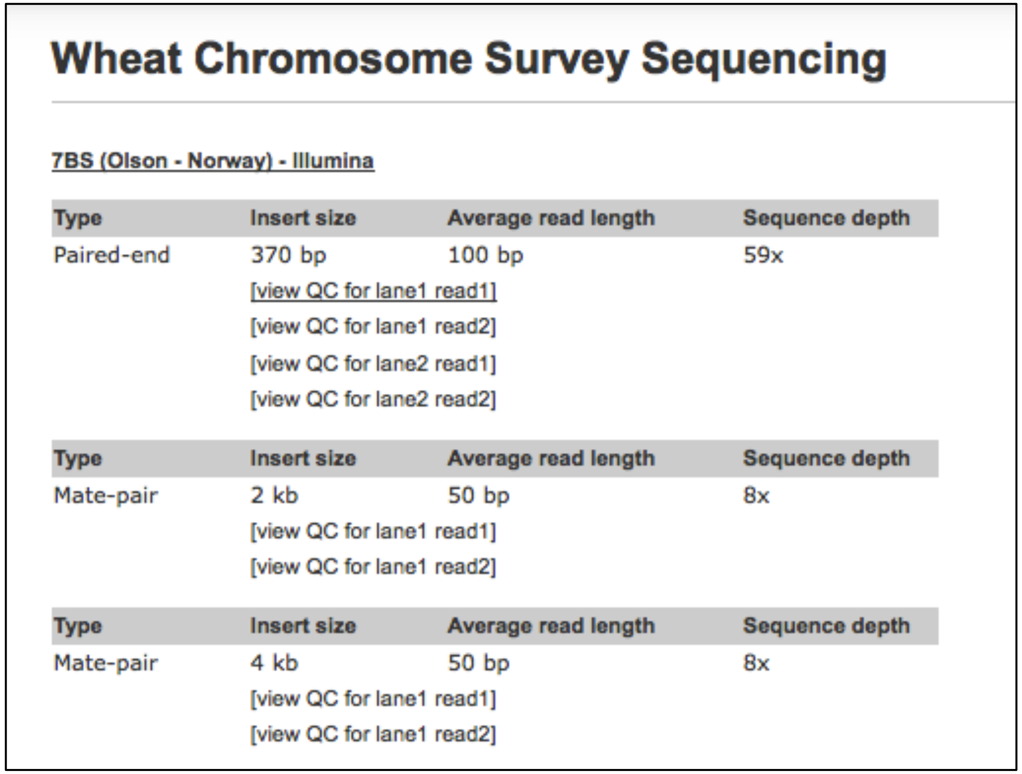
Tracking & Versioning

- **Data delivery:**
  - send data in hard drives as fastq sequence files but…
  - we are happy to assist with other formats and methods.

- **Report reception of data**

- **Summary file download (coming up)**

QC Checks



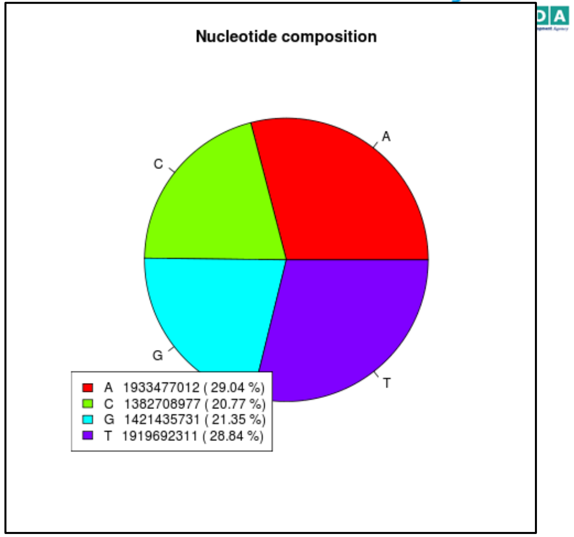
Wheat Chromosome Survey Sequencing

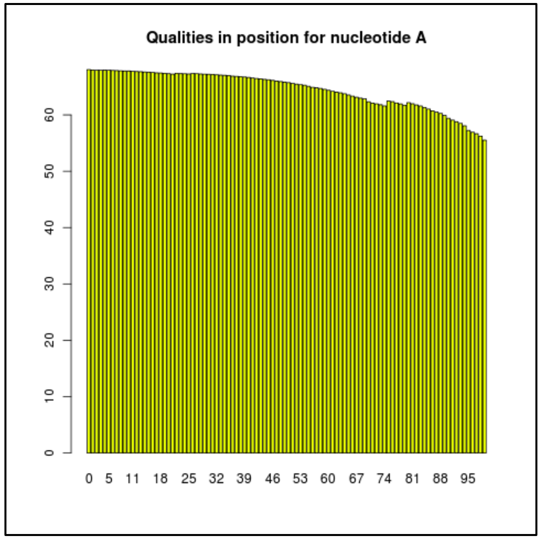
7BS (Olson - Norway) - Illumina

Type	Insert size	Average read length	Sequence depth
Paired-end	370 bp	100 bp	59x
	[view QC for lane1 read1]		
	[view QC for lane1 read2]		
	[view QC for lane2 read1]		
	[view QC for lane2 read2]		

Type	Insert size	Average read length	Sequence depth
Mate-pair	2 kb	50 bp	8x
	[view QC for lane1 read1]		
	[view QC for lane1 read2]		

Type	Insert size	Average read length	Sequence depth
Mate-pair	4 kb	50 bp	8x
	[view QC for lane1 read1]		
	[view QC for lane1 read2]		





Assembly Strategy

- **Assembly Tools**

  Newbler, Velvet, ABySS, Cortex, SGA, others

- **Parameters**

  K-mer size, coverage criteria, pair-ends, etc

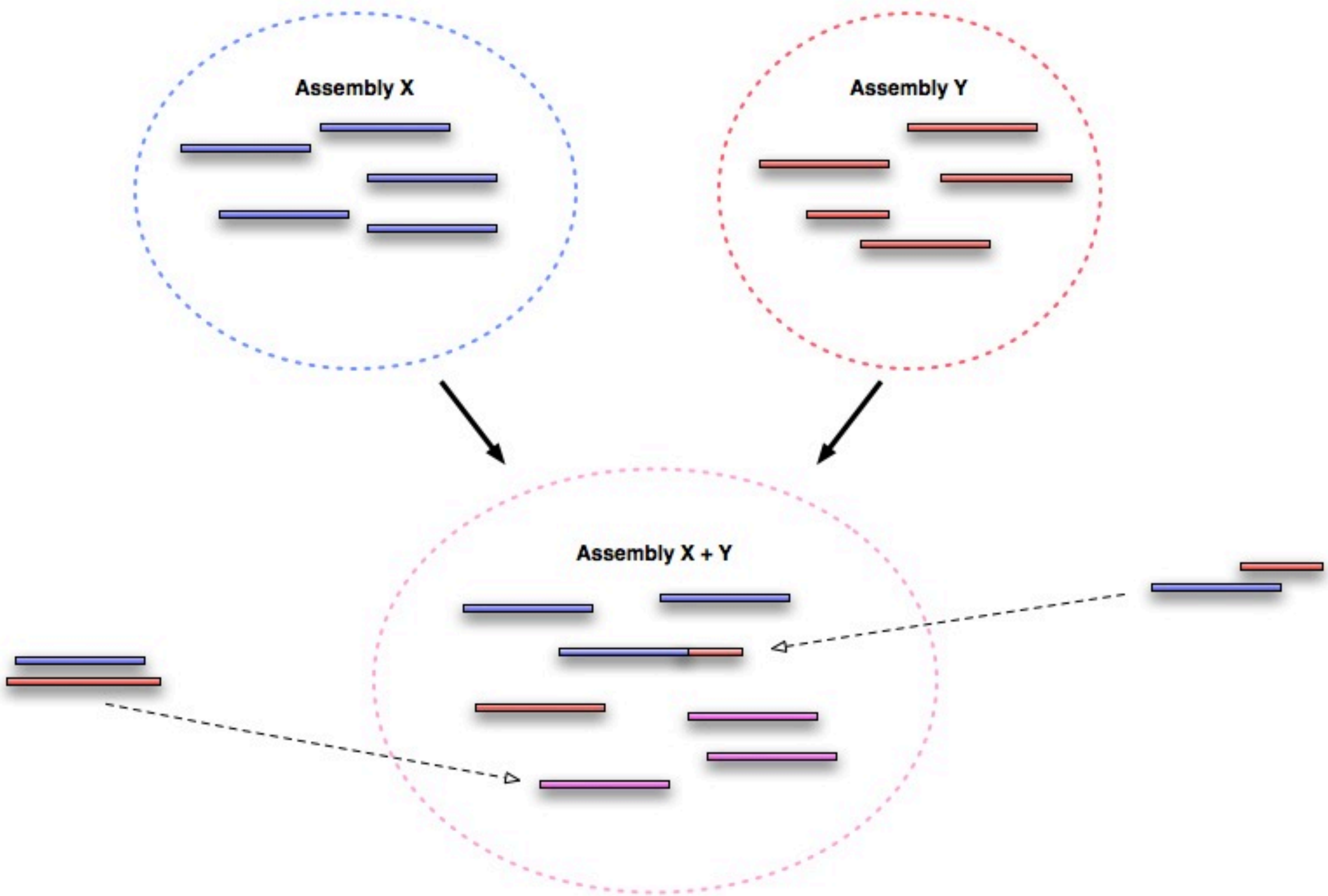
- **Evaluation**

  N50, number of contigs, screen for contamination

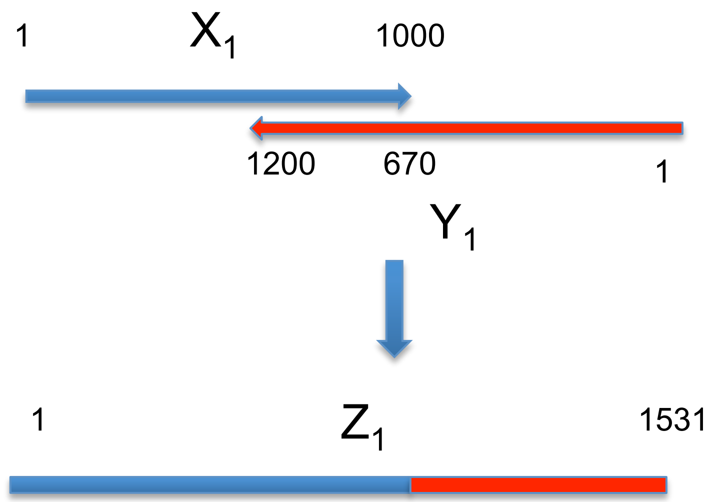
- **Assembly Consolidation**

  Aim: **“one assembly per chromosome arm”** per data freeze.

Assemblies Consolidation



Assemblies Consolidation



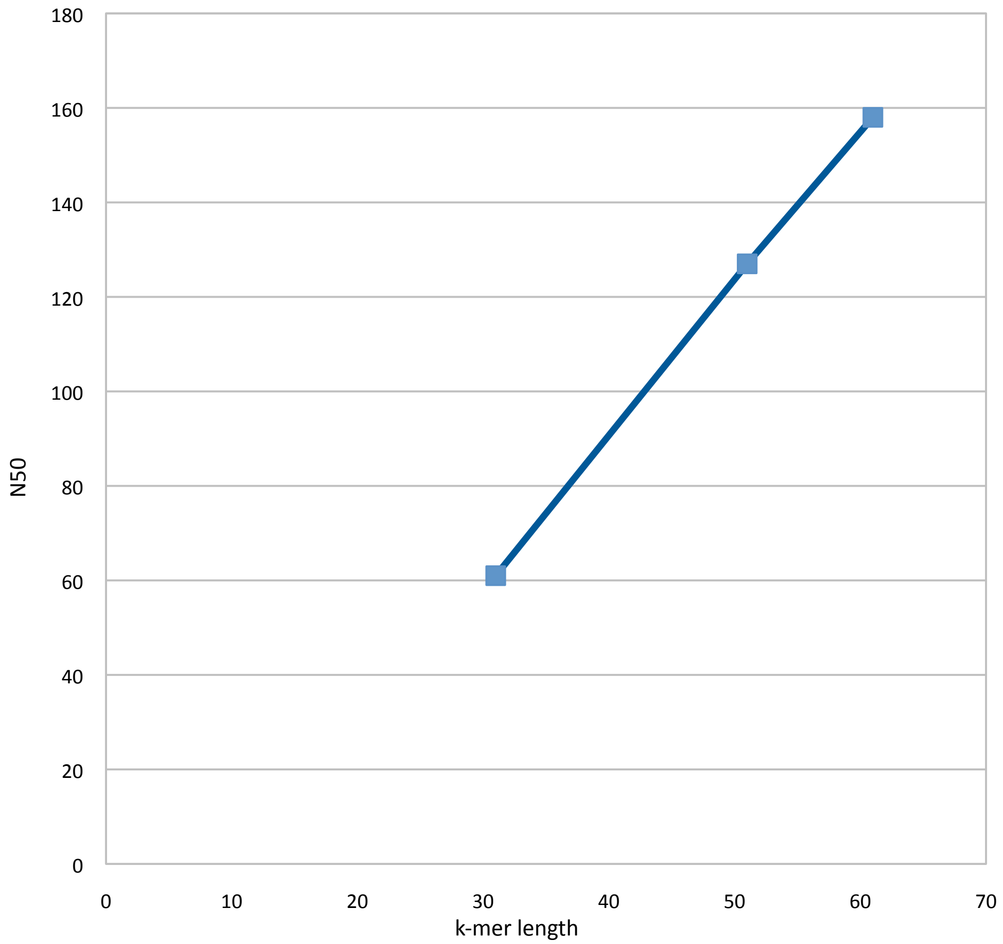
AGP

```
Z1    1      1000 1   W   X1   1     1000   +
Z1    1000   1531 2   W   Y1   1      670   -
```

www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/AGP_Specification.shtml

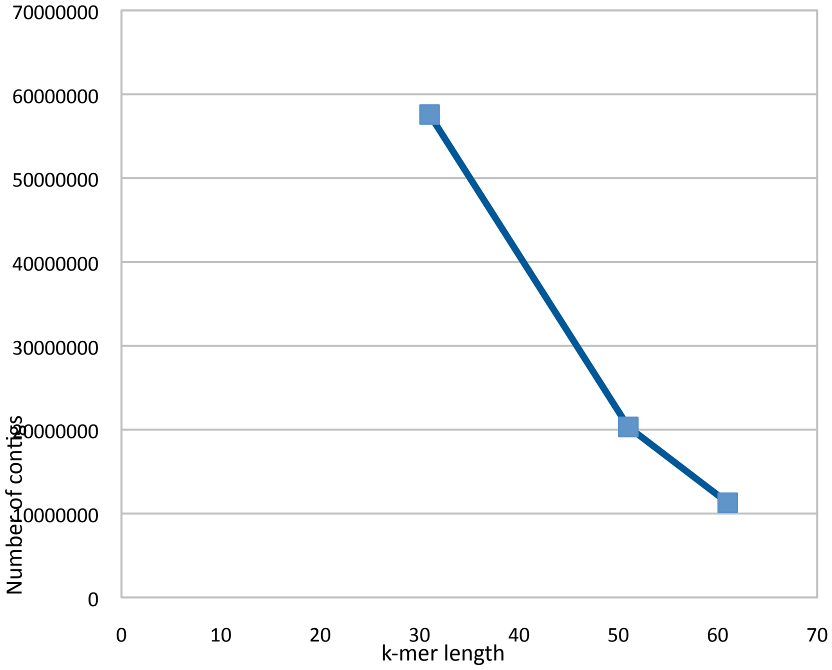
Preliminary Assemblies – 6BL

73x – Illumina – 100 bps - ~36Gb



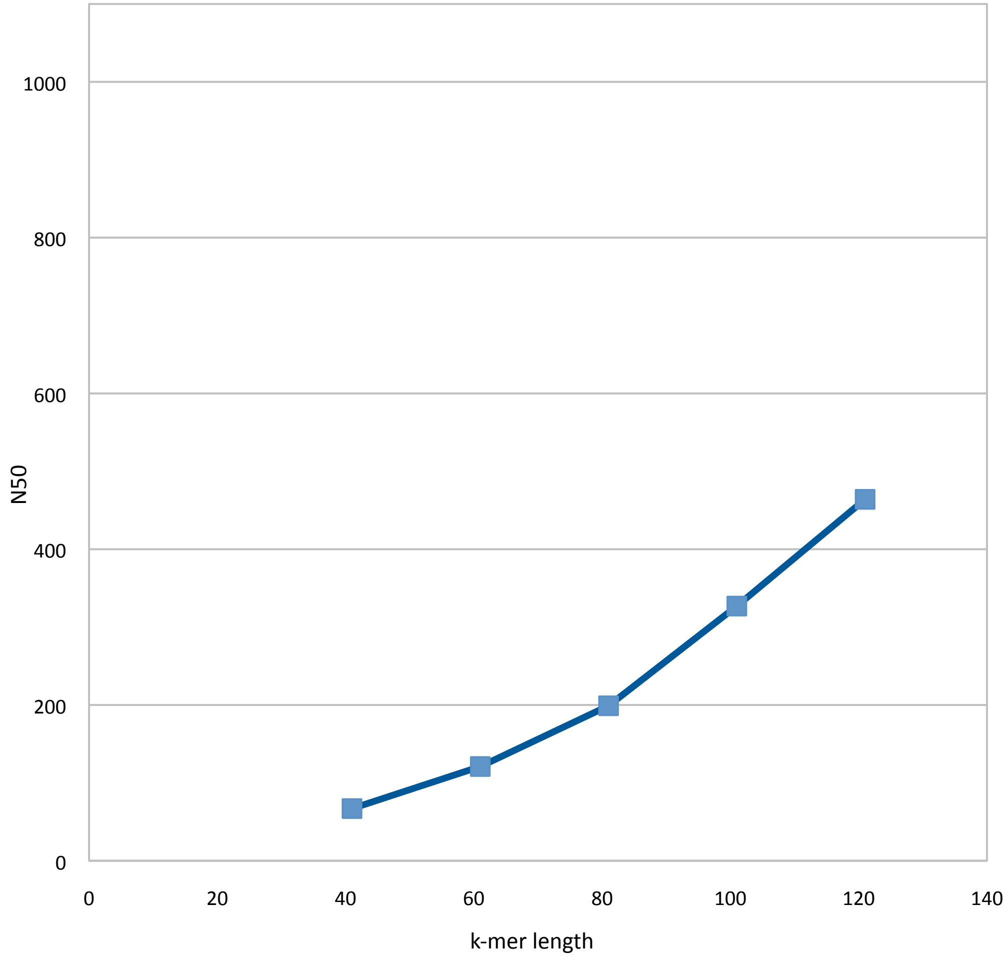
Preliminary Assemblies – 6BL

73x – Illumina – 100 bps - ~36Gb



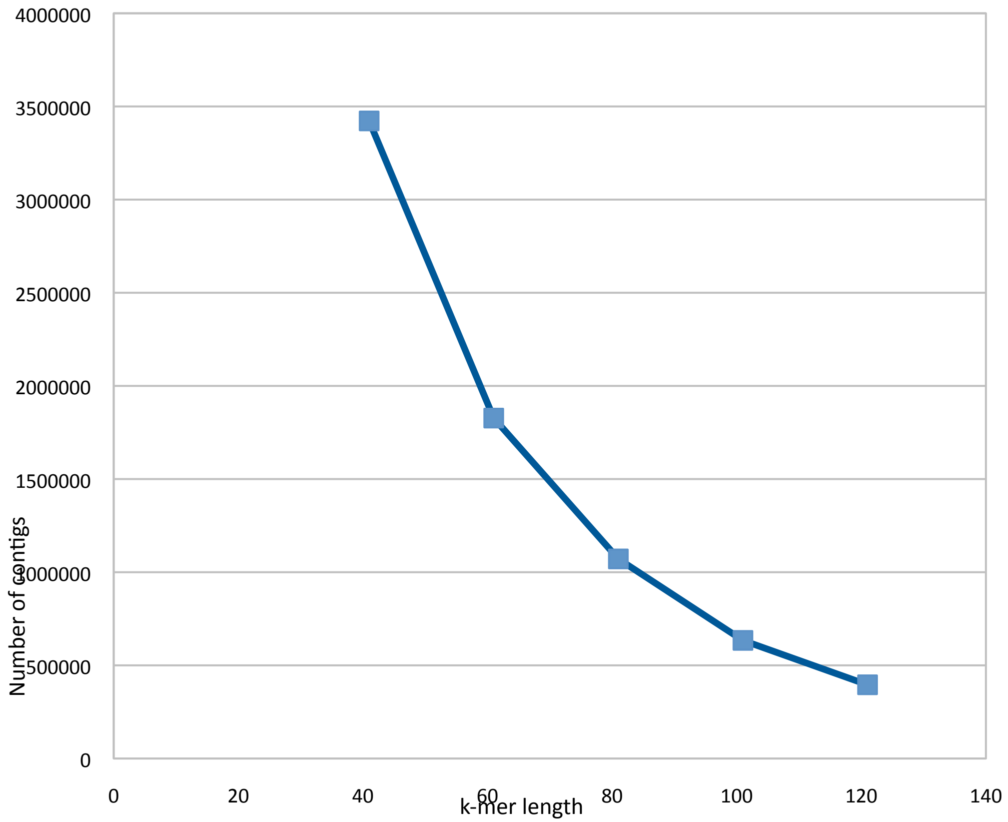
Preliminary Assemblies – 4DS

5x – 454 sequences



Preliminary Assemblies – 4DS

5x – 454 sequences



What can we do with the *Survey Sequences?*

- Annotate genes within contigs (intron-exon structure)

- Link features to chromosomes (within subgenomes)

- Localised synteny studies

- Approximate some of the global figures
  - Gene counts
  - Pesudogenes
  - Linage specific genes
  - Comparative analysis of homoeologous genes

What can’t we do with the *Survey Sequences?*

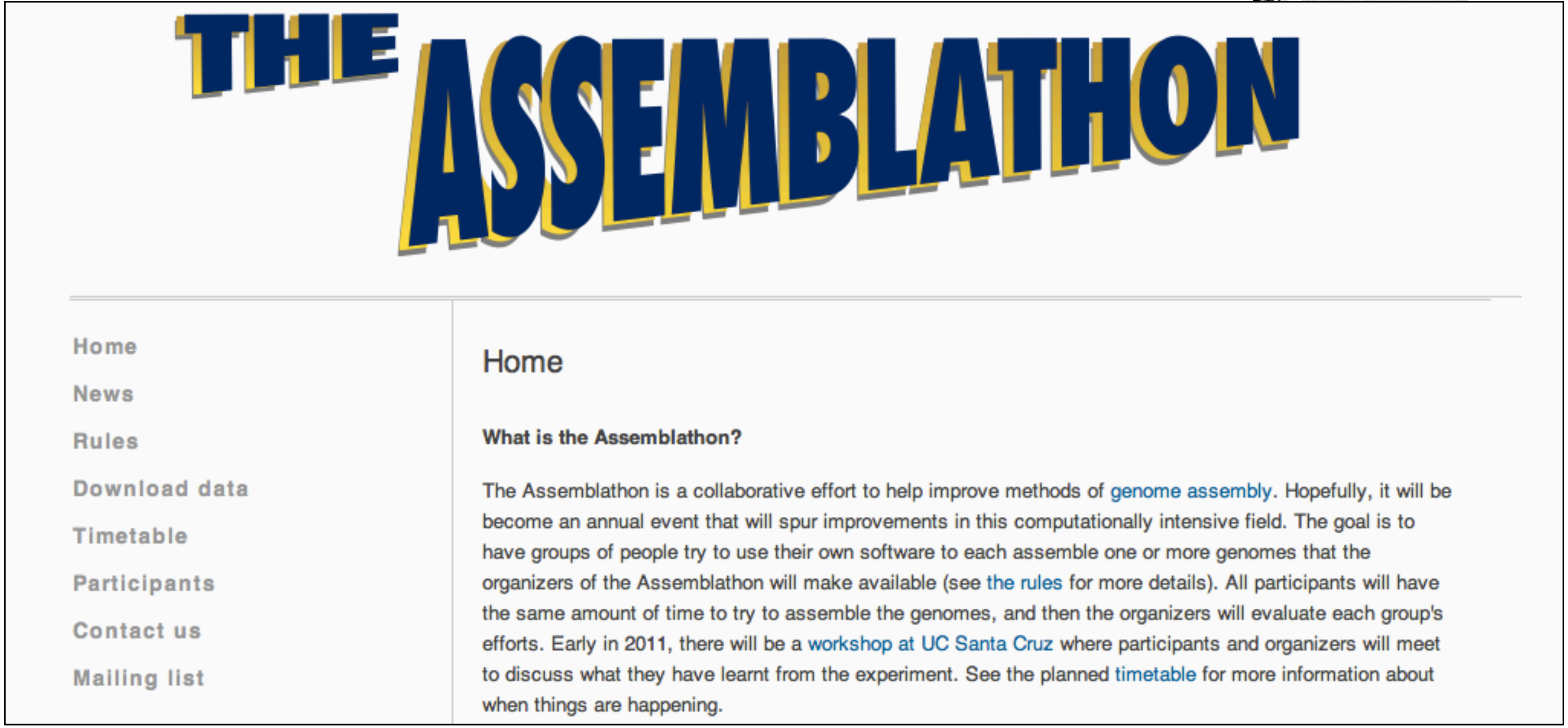
- It is not a going to give us a complete & finished genome
- Order and orientation of contigs will be only partial
- Global synteny studies comprising long contigs
- Re-arrangements will be difficult to detect
- Long range regulatory elements
- LD blocks…
- CNVs, structual variants ….

The Assemblathon – UC Santa Cruz

BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing



Home

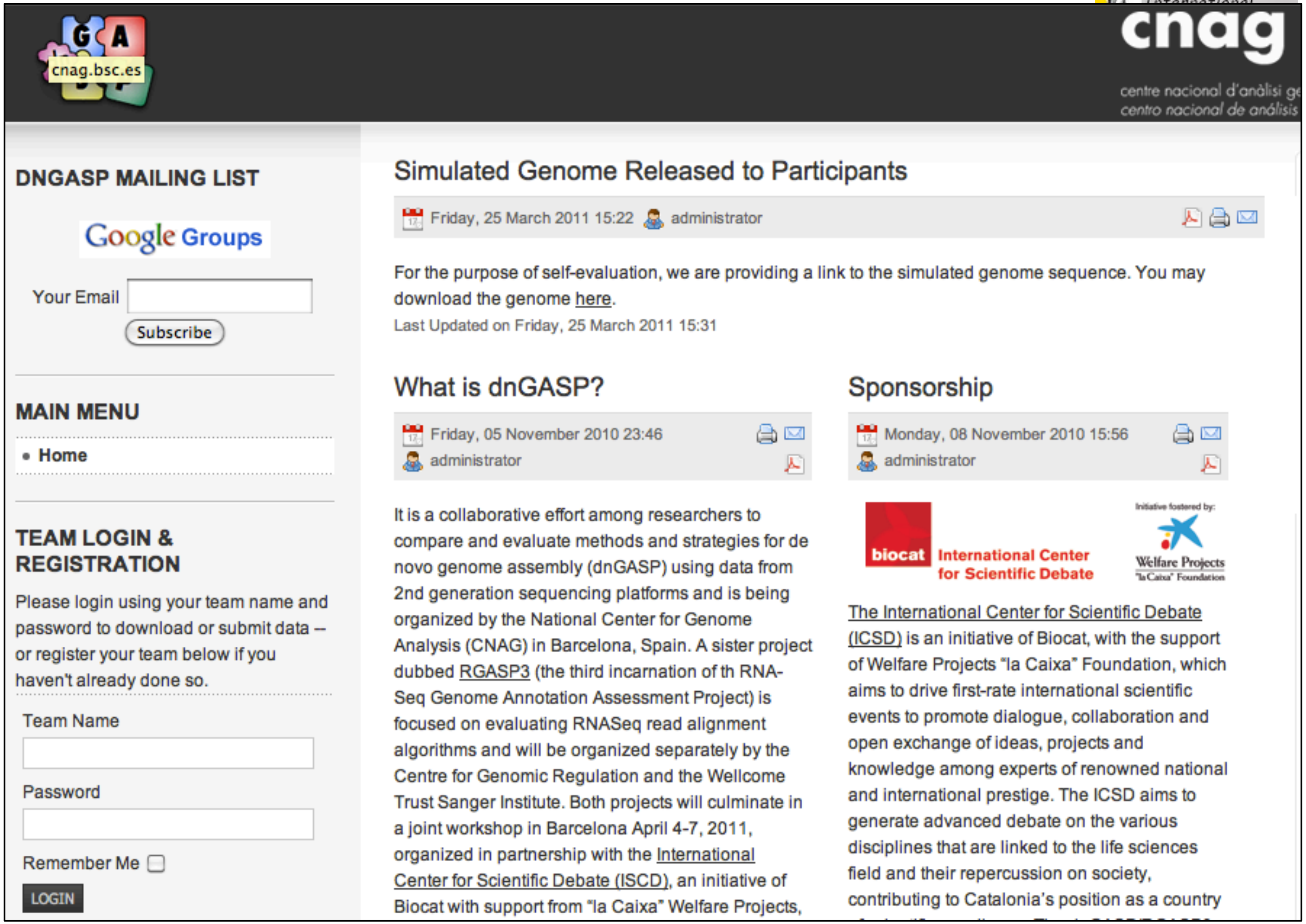
What is the Assemblathon?

The Assemblathon is a collaborative effort to help improve methods of genome assembly. Hopefully, it will be become an annual event that will spur improvements in this computationally intensive field. The goal is to have groups of people try to use their own software to each assemble one or more genomes that the organizers of the Assemblathon will make available (see the rules for more details). All participants will have the same amount of time to try to assemble the genomes, and then the organizers will evaluate each group's efforts. Early in 2011, there will be a workshop at UC Santa Cruz where participants and organizers will meet to discuss what they have learnt from the experiment. See the planned timetable for more information about when things are happening.

assemblathon.org

dnGASP

BBSRC Genome Analysis Centre
In partnership with EEDA Greater Norwich Development Partnership



cnag.bsc.es

cnag
centre nacional d'anàlisi ge
centro nacional de análisis

DNGASP MAILING LIST

Google Groups

Your Email

Subscribe

MAIN MENU

- Home

TEAM LOGIN & REGISTRATION

Please login using your team name and password to download or submit data – or register your team below if you haven't already done so.

Team Name

Password

Remember Me □

LOGIN

Simulated Genome Released to Participants

Friday, 25 March 2011 15:22 administrator

For the purpose of self-evaluation, we are providing a link to the simulated genome sequence. You may download the genome here.

Last Updated on Friday, 25 March 2011 15:31

What is dnGASP?

Friday, 05 November 2010 23:46
administrator

It is a collaborative effort among researchers to compare and evaluate methods and strategies for de novo genome assembly (dnGASP) using data from 2nd generation sequencing platforms and is being organized by the National Center for Genome Analysis (CNAG) in Barcelona, Spain. A sister project dubbed RGASP3 (the third incarnation of th RNA-Seq Genome Annotation Assessment Project) is focused on evaluating RNASeq read alignment algorithms and will be organized separately by the Centre for Genomic Regulation and the Wellcome Trust Sanger Institute. Both projects will culminate in a joint workshop in Barcelona April 4-7, 2011, organized in partnership with the International Center for Scientific Debate (ISCD), an initiative of Biocat with support from "la Caixa" Welfare Projects,

Sponsorship

Monday, 08 November 2010 15:56
administrator

biocat International Center for Scientific Debate

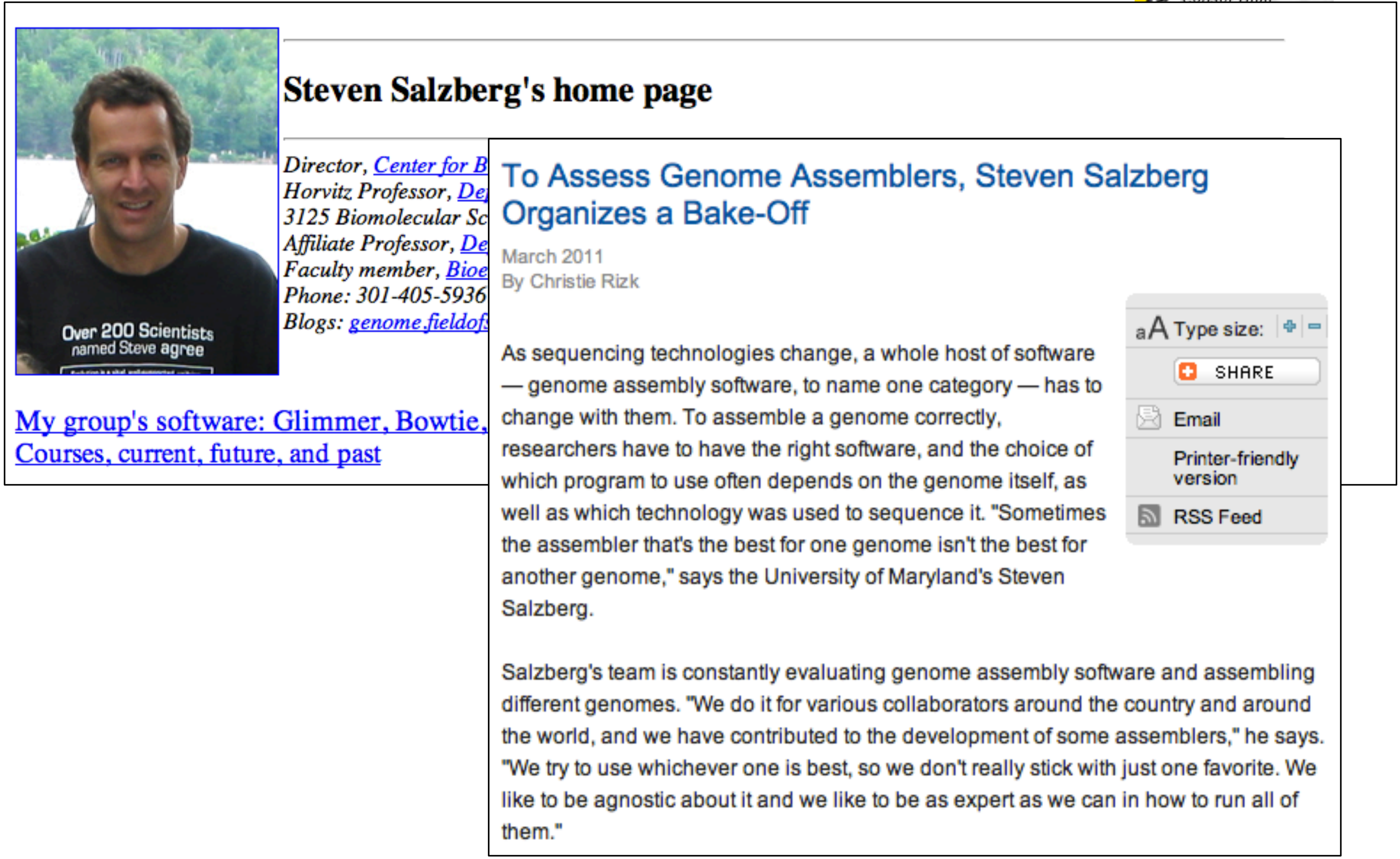
Initiative fostered by: Welfare Projects "la Caixa" Foundation

The International Center for Scientific Debate (ICSD) is an initiative of Biocat, with the support of Welfare Projects "la Caixa" Foundation, which aims to drive first-rate international scientific events to promote dialogue, collaboration and open exchange of ideas, projects and knowledge among experts of renowned national and international prestige. The ICSD aims to generate advanced debate on the various disciplines that are linked to the life sciences field and their repercussion on society, contributing to Catalonia's position as a country

Salzberg’s bakeoff

BBSRC Genome Analysis Centre
In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium



Steven Salzberg's home page

Director, Center for B
Horvitz Professor, De
3125 Biomolecular Sc
Affiliate Professor, De
Faculty member, Bioe
Phone: 301-405-5936
Blogs: genome.fieldof

My group's software: Glimmer, Bowtie,
Courses, current, future, and past

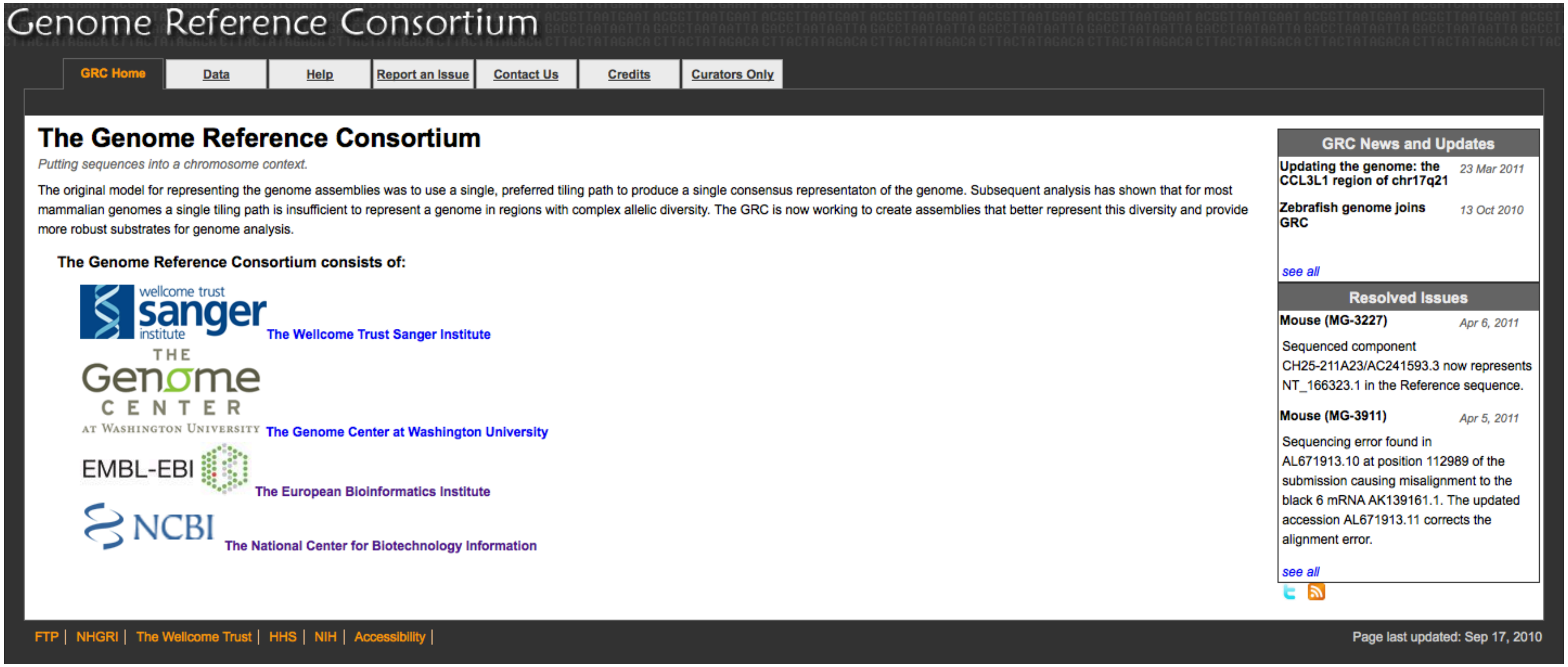
To Assess Genome Assemblers, Steven Salzberg Organizes a Bake-Off

March 2011
By Christie Rizk

As sequencing technologies change, a whole host of software — genome assembly software, to name one category — has to change with them. To assemble a genome correctly, researchers have to have the right software, and the choice of which program to use often depends on the genome itself, as well as which technology was used to sequence it. "Sometimes the assembler that's the best for one genome isn't the best for another genome," says the University of Maryland's Steven Salzberg.

Salzberg's team is constantly evaluating genome assembly software and assembling different genomes. "We do it for various collaborators around the country and around the world, and we have contributed to the development of some assemblers," he says. "We try to use whichever one is best, so we don't really stick with just one favorite. We like to be agnostic about it and we like to be as expert as we can in how to run all of them."

What is next?



Genome Reference Consortium

The Genome Reference Consortium

Putting sequences into a chromosome context.

The original model for representing the genome assemblies was to use a single, preferred tiling path to produce a single consensus representaton of the genome. Subsequent analysis has shown that for most mammalian genomes a single tiling path is insufficient to represent a genome in regions with complex allelic diversity. The GRC is now working to create assemblies that better represent this diversity and provide more robust substrates for genome analysis.

The Genome Reference Consortium consists of:

- The Wellcome Trust Sanger Institute
- The Genome Center at Washington University
- The European Bioinformatics Institute
- The National Center for Biotechnology Information

GRC News and Updates

Updating the genome: the CCL3L1 region of chr17q21	*23 Mar 2011*
Zebrafish genome joins GRC	*13 Oct 2010*

see all

Resolved Issues

Mouse (MG-3227) *Apr 6, 2011*

Sequenced component CH25-211A23/AC241593.3 now represents NT_166323.1 in the Reference sequence.

Mouse (MG-3911) *Apr 5, 2011*

Sequencing error found in AL671913.10 at position 112989 of the submission causing misalignment to the black 6 mRNA AK139161.1. The updated accession AL671913.11 corrects the alignment error.

see all

FTP | NHGRI | The Wellcome Trust | HHS | NIH | Accessibility |

Page last updated: Sep 17, 2010

Agenda

- Wheat Chromosome Sequencing Survey DCC
- Assemblies - theory
- Assemblies - practice

Assemblies

- **The problem**

  "Assembly for Large Genomes"

- **The solutions**

  Overlap Graphs

  De Bruijn Graphs

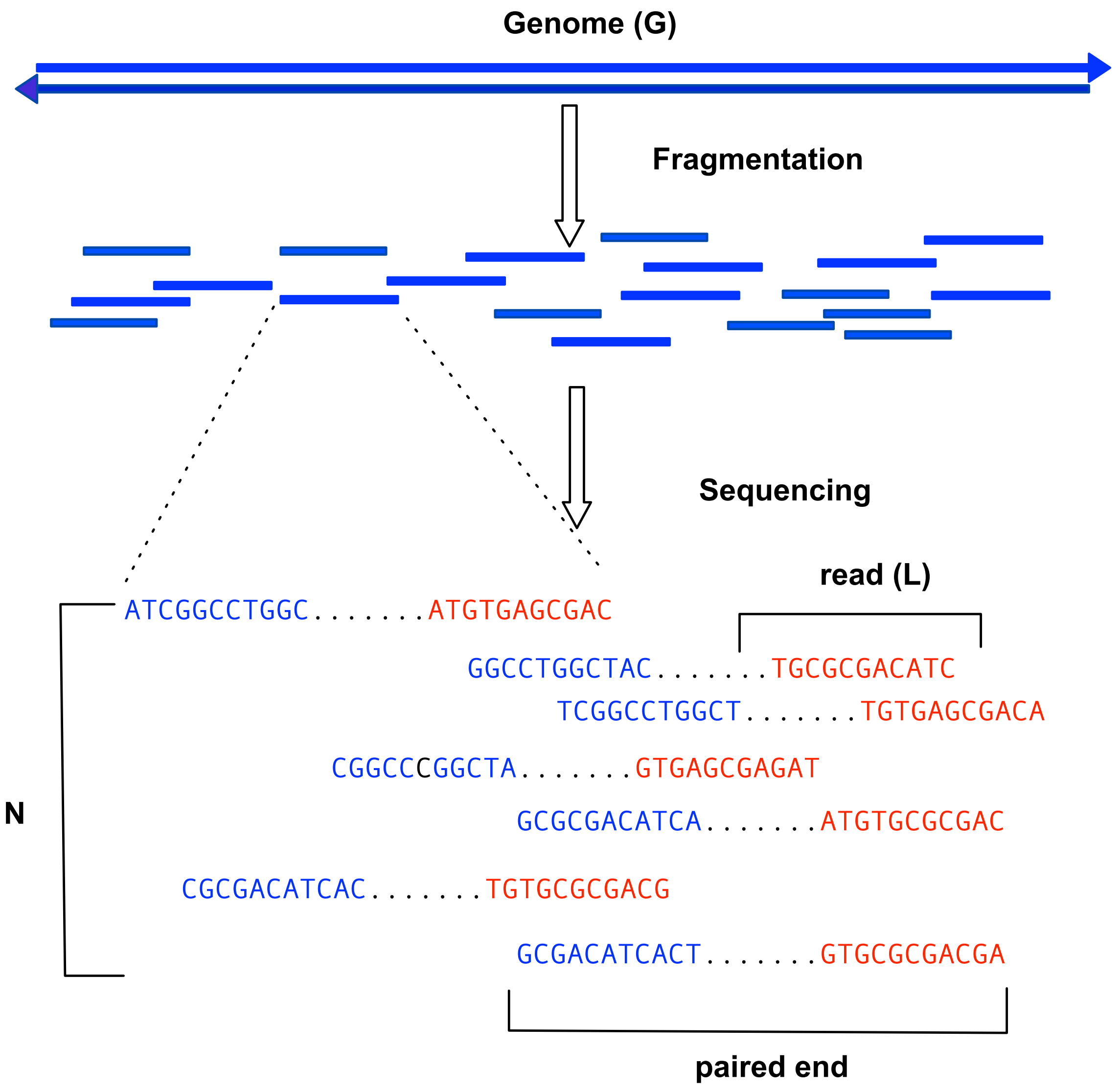
  String Graphs

- **The challenges**

  1. Far too many reads
  2. Lack of coverage
  3. Memory-hungry algorithms
  4. Sequencing error profiles

The Assembly Problem

Sequencing a Genome



ATCGGCCTGGC.......ATGTGAGCGAC

GGCCTGGCTAC.......TGCGCGACATC

TCGGCCTGGCT.......TGTGAGCGACA

CGGCCCGGCTA.......GTGAGCGAGAT

GCGCGACATCA.......ATGTGCGCGAC

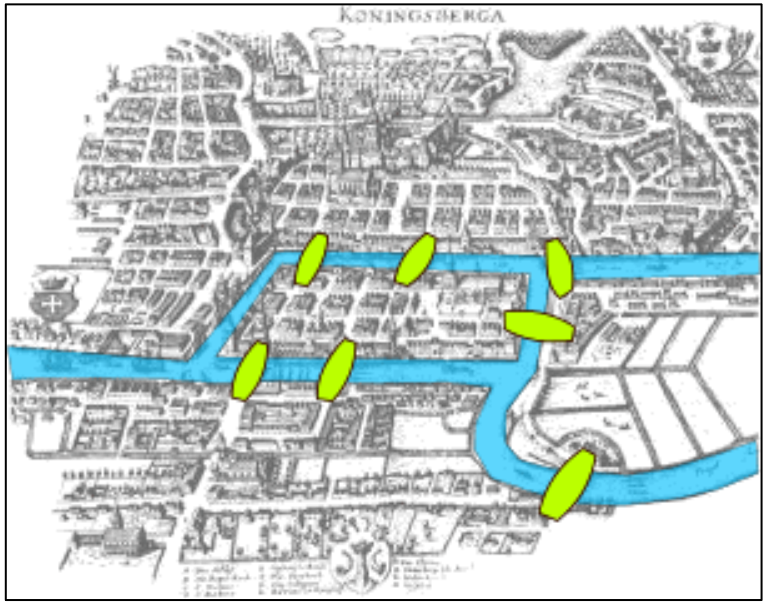
CGCGACATCAC.......TGTGCGCGACG

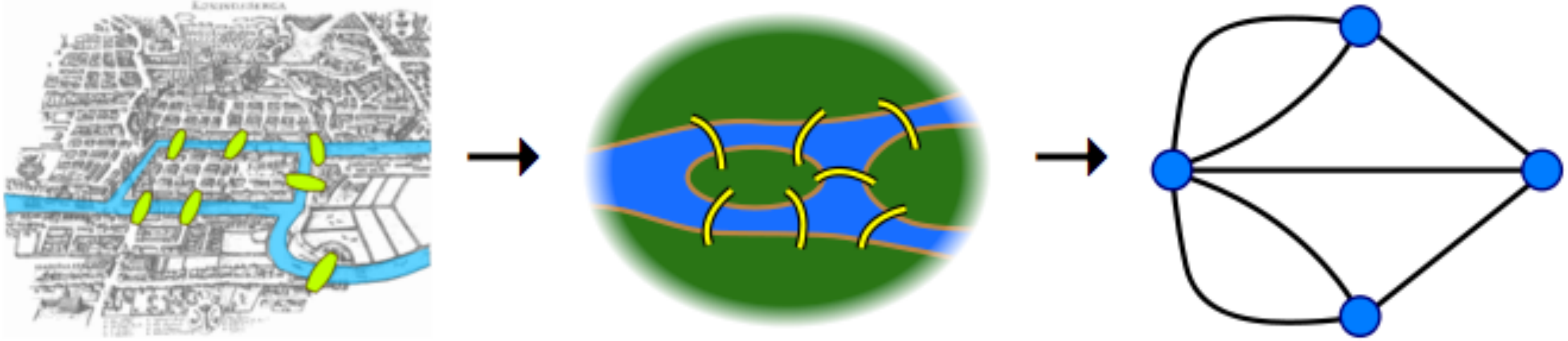
GCGACATCACT.......GTGCGCGACGA

Graph Theory

Leonhard Euler (1707-1783)

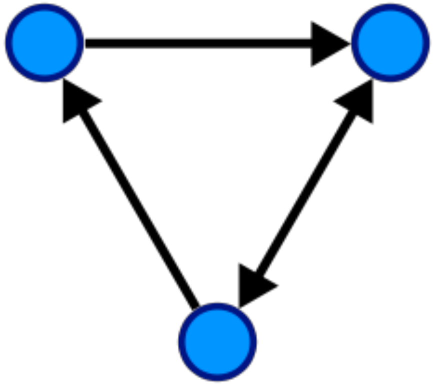
Seven Bridges of Königsberg

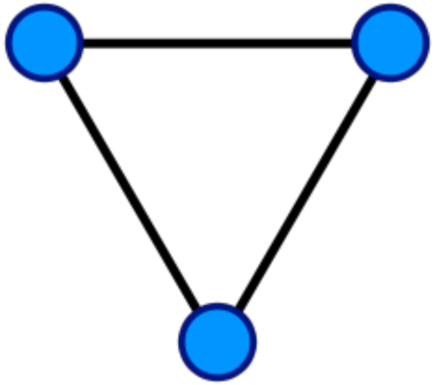


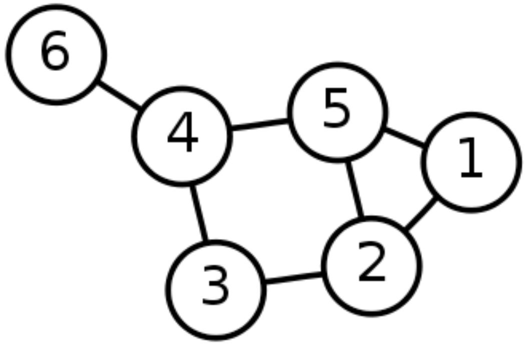


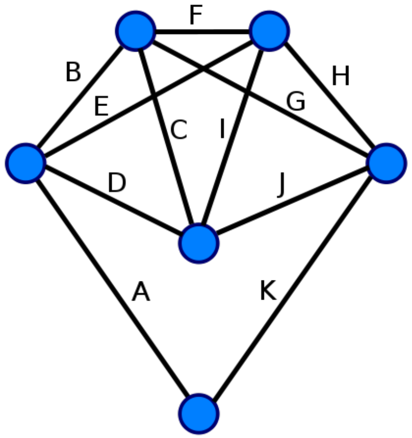
Graphs

nodes & edges



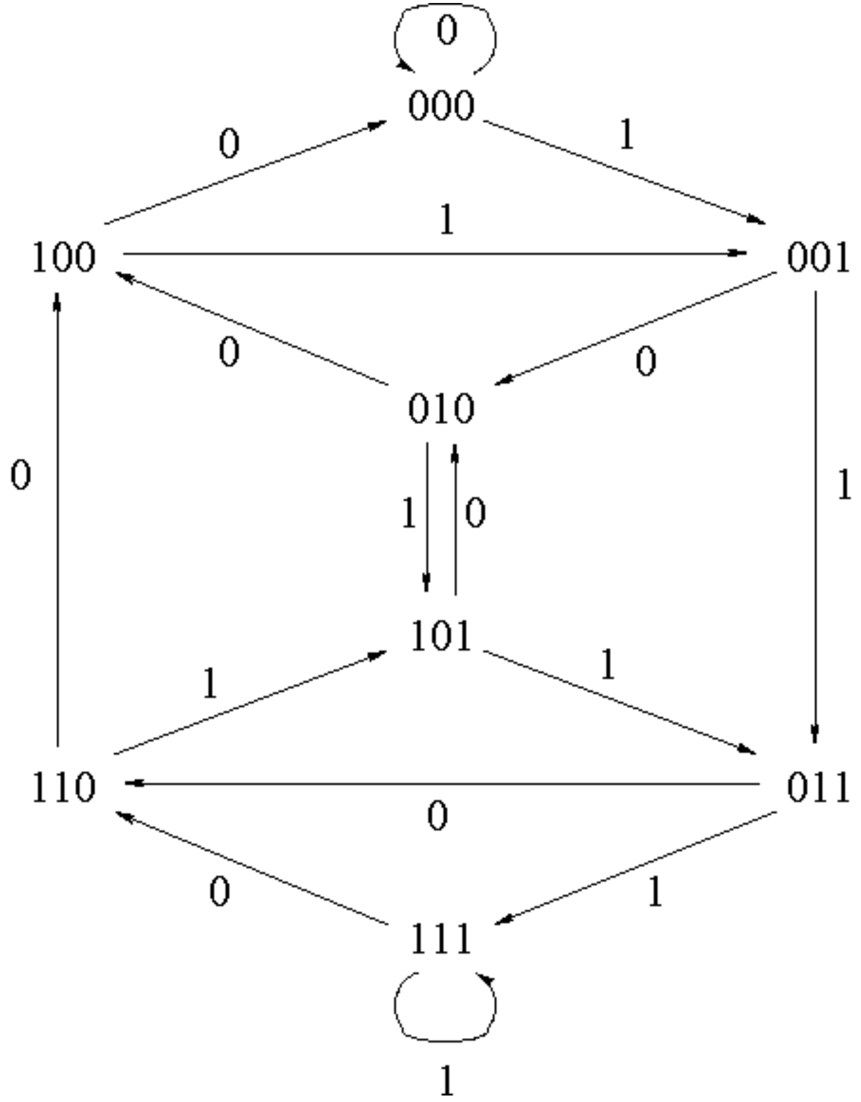




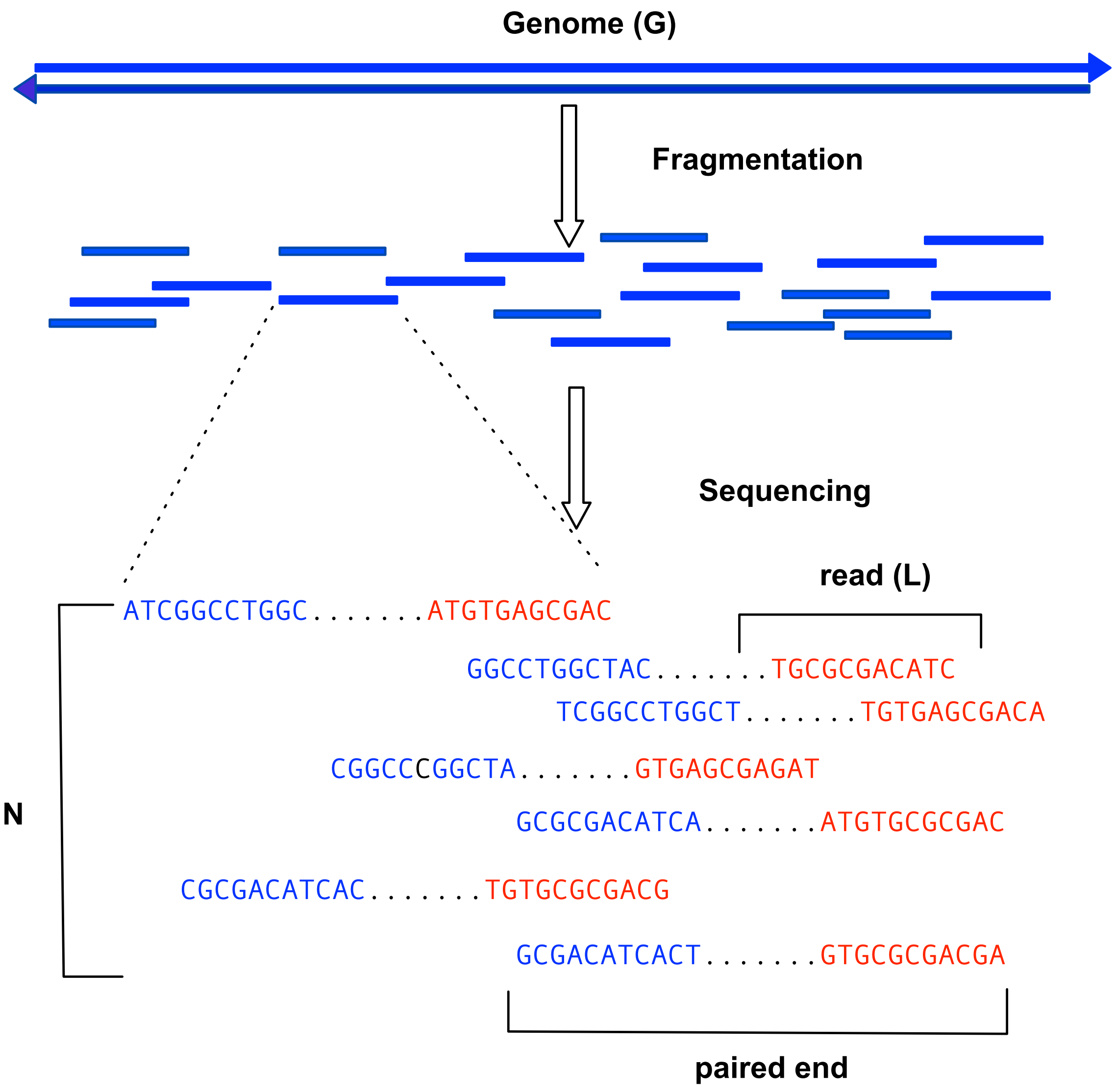


Walk in the graph

Eulerian paths versus Hamiltonian paths



Sequencing a Genome



Genome (G)

Fragmentation

Sequencing

read (L)

ATCGGCCTGGC.......ATGTGAGCGAC

GGCCTGGCTAC.......TGCGCGACATC

TCGGCCTGGCT.......TGTGAGCGACA

CGGCCCGGCTA.......GTGAGCGAGAT

GCGCGACATCA.......ATGTGCGCGAC

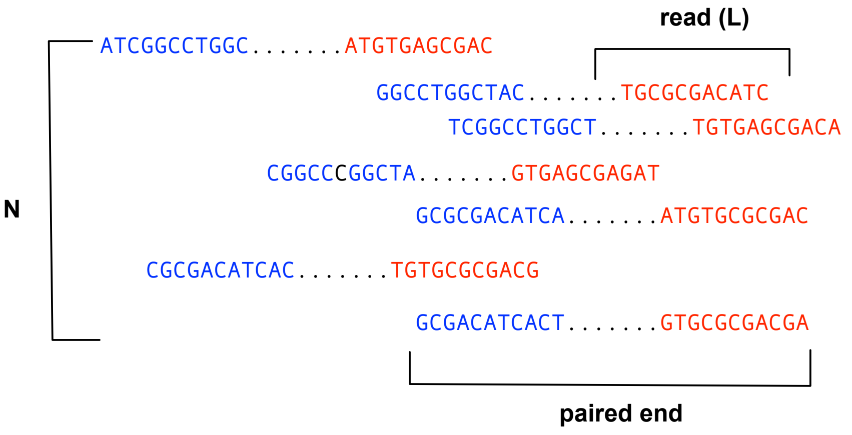
CGCGACATCAC.......TGTGCGCGACG

GCGACATCACT.......GTGCGCGACGA

N

paired end

The Assembly Problem



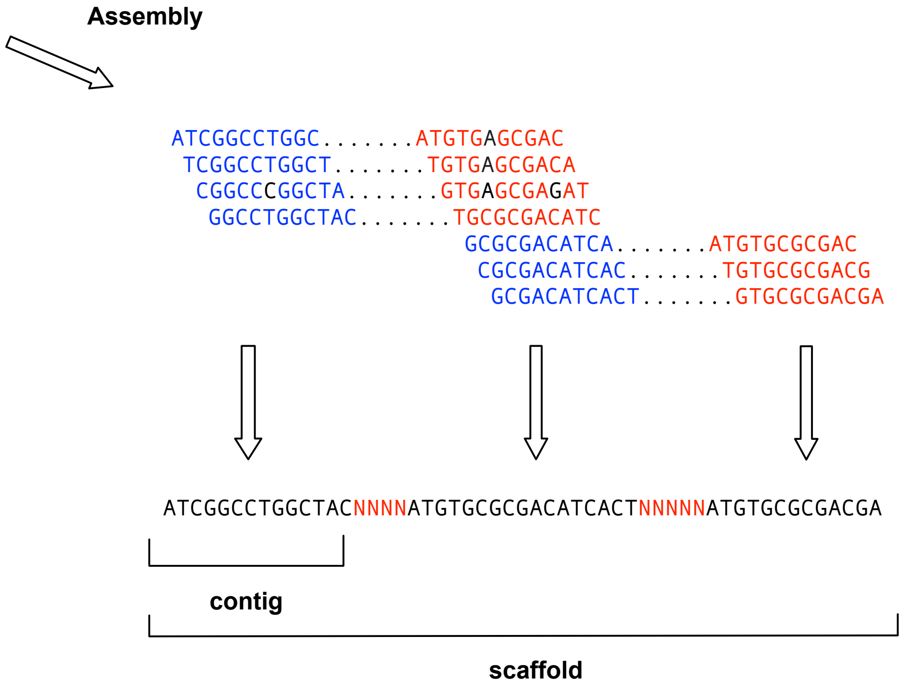
read (L)

N

```
ATCGGCCTGGC.......ATGTGAGCGAC
                         GGCCTGGCTAC.......TGCGCGACATC
                               TCGGCCTGGCT.......TGTGAGCGACA
               CGGCCCGGCTA.......GTGAGCGAGAT
                            GCGCGACATCA.......ATGTGCGCGAC
    CGCGACATCAC.......TGTGCGCGACG
                            GCGACATCACT.......GTGCGCGACGA
```

paired end

Coverage: (N * L) /G



Assembly

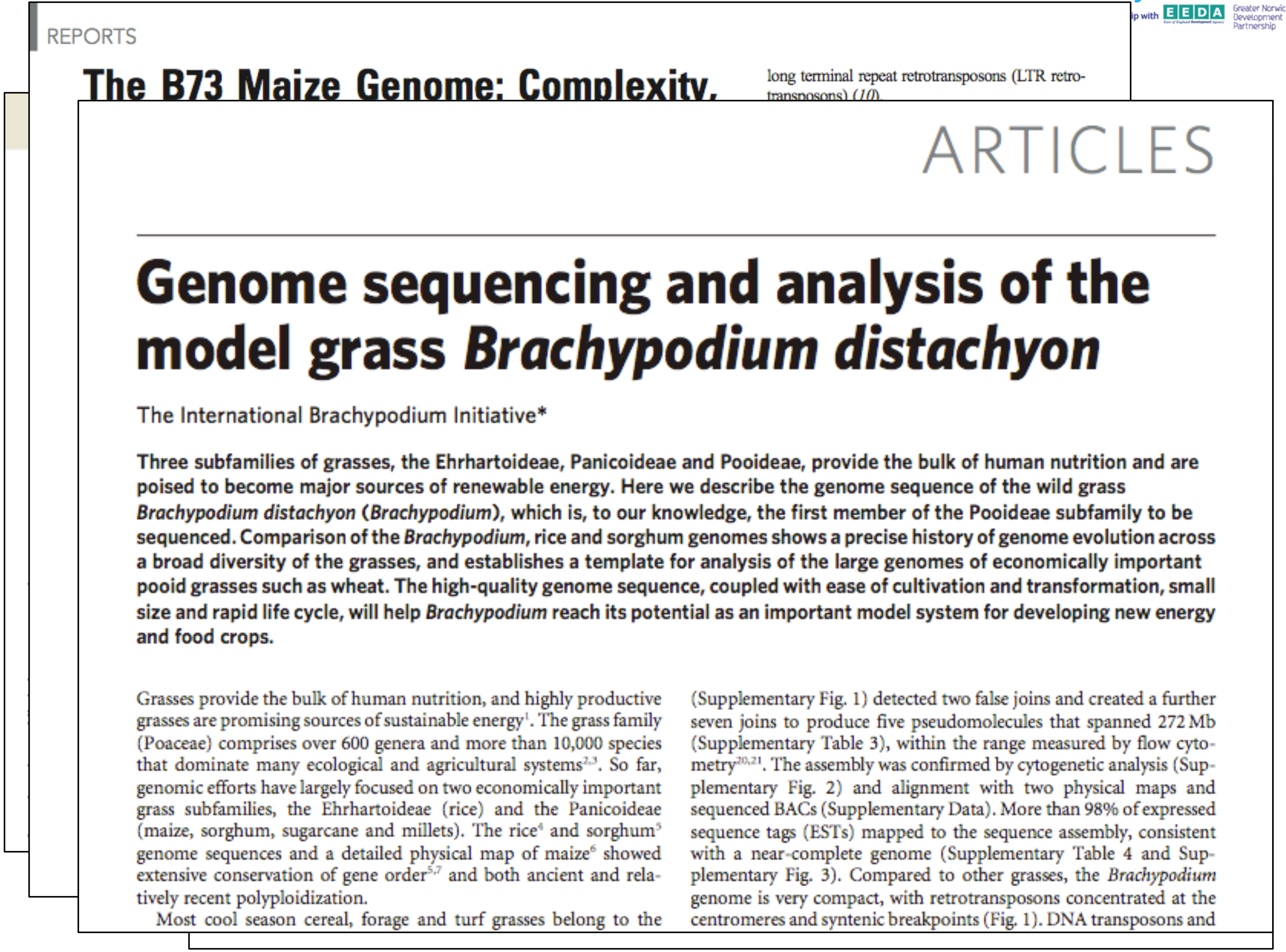
```
ATCGGCCTGGC.......ATGTGAGCGAC
 TCGGCCTGGCT.......TGTGAGCGACA
  CGGCCCGGCTA.......GTGAGCGAGAT
   GGCCTGGCTAC.......TGCGCGACATC
                      GCGCGACATCA.......ATGTGCGCGAC
                       CGCGACATCAC.......TGTGCGCGACG
                        GCGACATCACT.......GTGCGCGACGA
```

```
ATCGGCCTGGCTACNNNNATGTGCGCGACATCACTNNNNNATGTGCGCGACGA
```

contig

scaffold

BBSRC Genome Analysis Centre
In partnership with EEDA Greater Norwich Development Partnership



REPORTS

The B73 Maize Genome: Complexity,

long terminal repeat retrotransposons (LTR retrotransposons) (10).

ARTICLES

Genome sequencing and analysis of the model grass *Brachypodium distachyon*

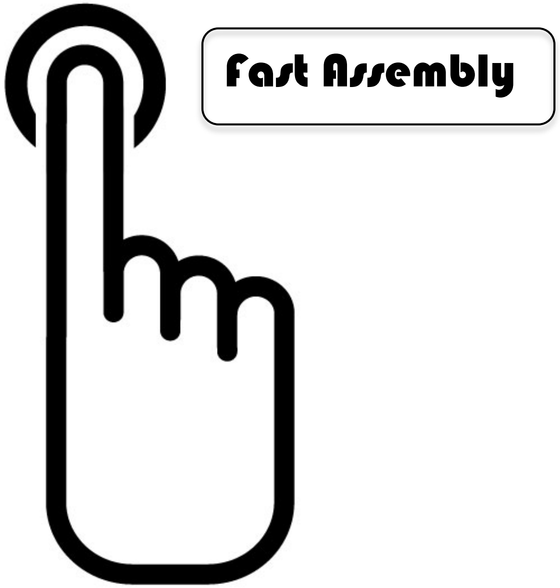
The International Brachypodium Initiative*

Three subfamilies of grasses, the Ehrhartoideae, Panicoideae and Pooideae, provide the bulk of human nutrition and are poised to become major sources of renewable energy. Here we describe the genome sequence of the wild grass *Brachypodium distachyon* (*Brachypodium*), which is, to our knowledge, the first member of the Pooideae subfamily to be sequenced. Comparison of the *Brachypodium*, rice and sorghum genomes shows a precise history of genome evolution across a broad diversity of the grasses, and establishes a template for analysis of the large genomes of economically important pooid grasses such as wheat. The high-quality genome sequence, coupled with ease of cultivation and transformation, small size and rapid life cycle, will help *Brachypodium* reach its potential as an important model system for developing new energy and food crops.

Grasses provide the bulk of human nutrition, and highly productive grasses are promising sources of sustainable energy[1]. The grass family (Poaceae) comprises over 600 genera and more than 10,000 species that dominate many ecological and agricultural systems[2,3]. So far, genomic efforts have largely focused on two economically important grass subfamilies, the Ehrhartoideae (rice) and the Panicoideae (maize, sorghum, sugarcane and millets). The rice[4] and sorghum[5] genome sequences and a detailed physical map of maize[6] showed extensive conservation of gene order[5,7] and both ancient and relatively recent polyploidization.

Most cool season cereal, forage and turf grasses belong to the

(Supplementary Fig. 1) detected two false joins and created a further seven joins to produce five pseudomolecules that spanned 272 Mb (Supplementary Table 3), within the range measured by flow cytometry[20,21]. The assembly was confirmed by cytogenetic analysis (Supplementary Fig. 2) and alignment with two physical maps and sequenced BACs (Supplementary Data). More than 98% of expressed sequence tags (ESTs) mapped to the sequence assembly, consistent with a near-complete genome (Supplementary Table 4 and Supplementary Fig. 3). Compared to other grasses, the *Brachypodium* genome is very compact, with retrotransposons concentrated at the centromeres and syntenic breakpoints (Fig. 1). DNA transposons and

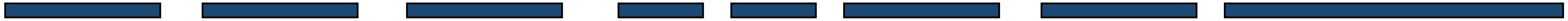


…. but we should approach an assembly as a lab experiment.

What is a good assembly?

- **Contiguity**
  - longest contig vs number of contigs
  - N50
- **Completeness**
  - gene count
  - gene coverage
- **Accuracy**
  - misassemblies (chimeric contigs)
  - base calls

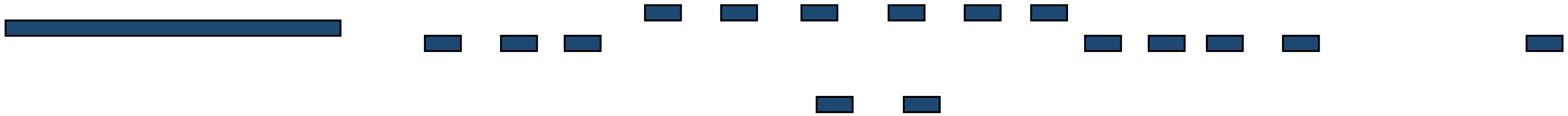
A good assembly?



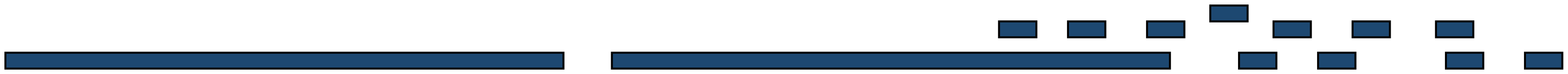
Largest contigs? Number of contigs?

A good assembly?

Largest contigs



Number of contigs

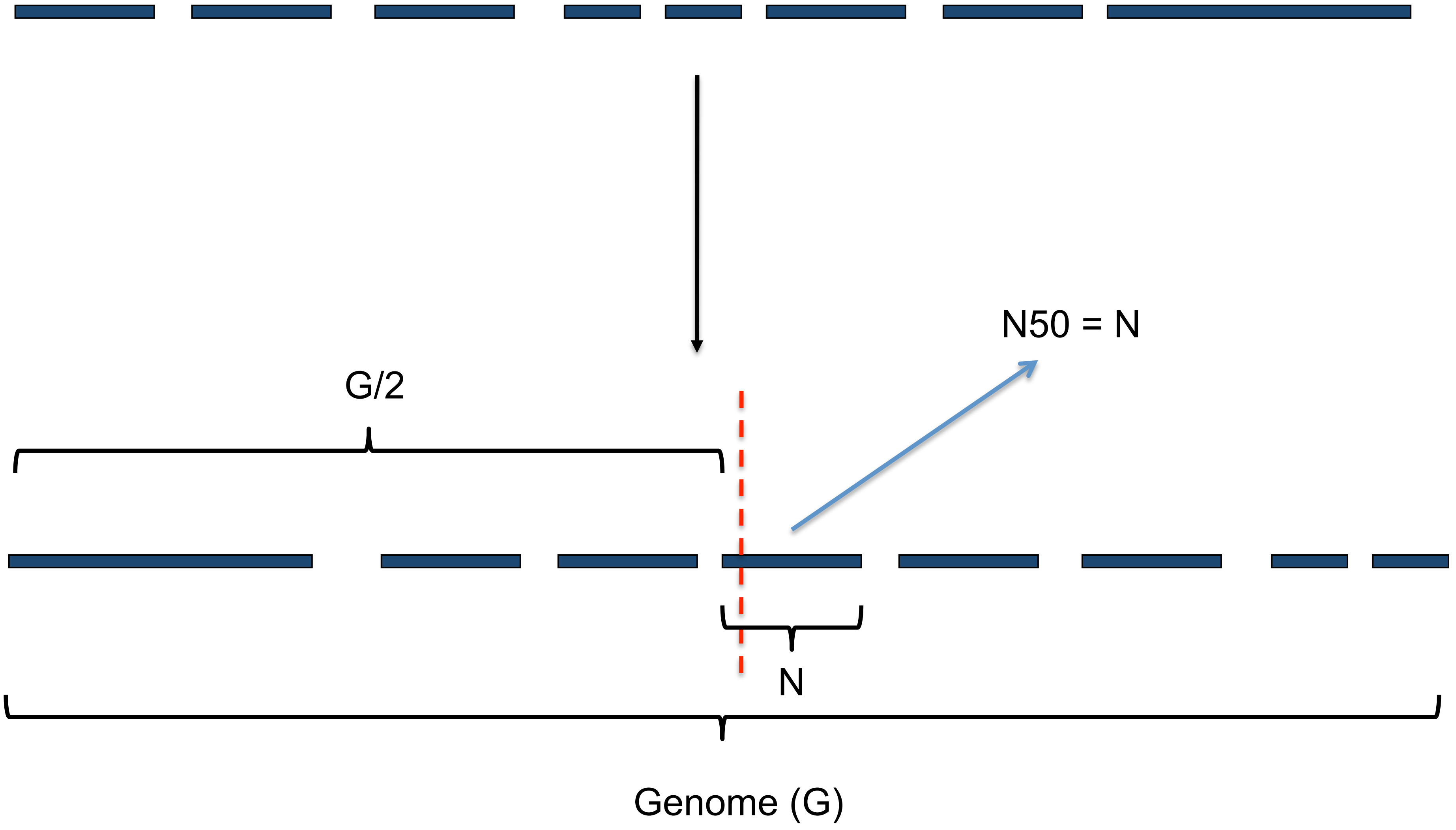


N50

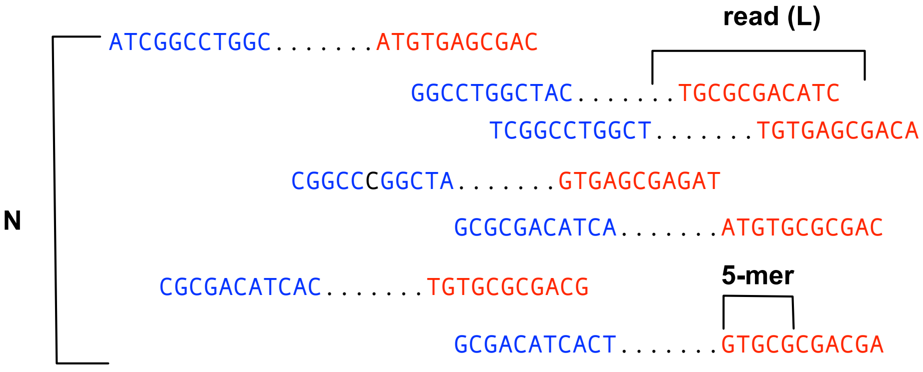
BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium



Overlap Graphs



read (L)

N

ATCGGCCTGGC.......ATGTGAGCGAC

GGCCTGGCTAC.......TGCGCGACATC

TCGGCCTGGCT.......TGTGAGCGACA

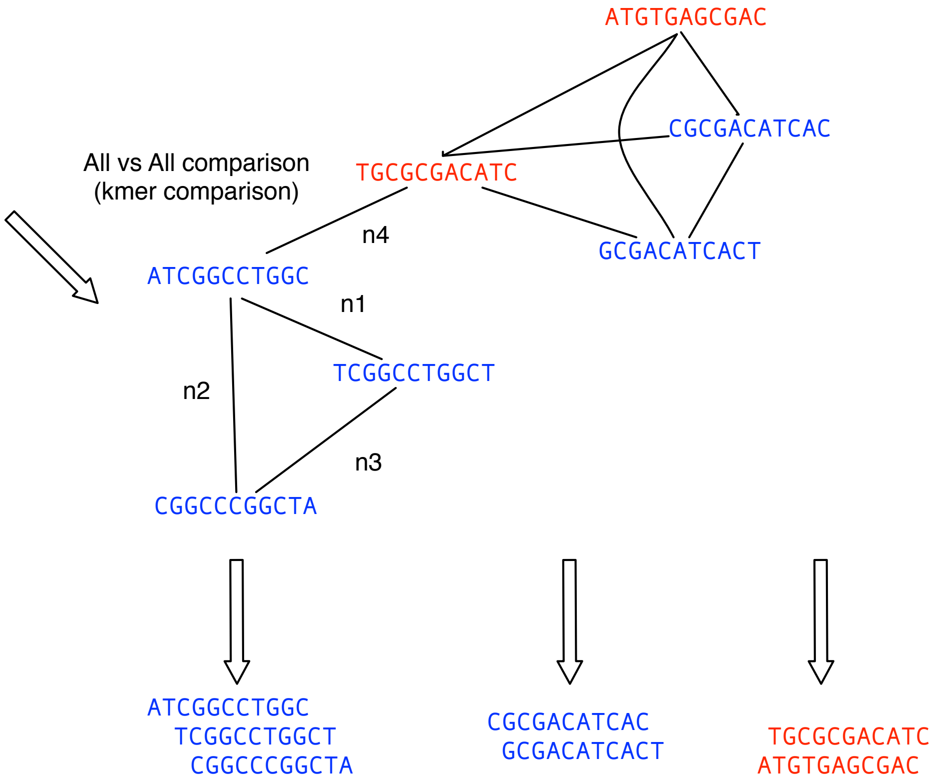
CGGCCCGGCTA.......GTGAGCGAGAT

GCGCGACATCA.......ATGTGCGCGAC

CGCGACATCAC.......TGTGCGCGACG

5-mer

GCGACATCACT.......GTGCGCGACGA



All vs All comparison
(kmer comparison)

ATGTGAGCGAC

CGCGACATCAC

TGCGCGACATC

GCGACATCACT

n4

ATCGGCCTGGC

n1

TCGGCCTGGCT

n2

n3

CGGCCCGGCTA

ATCGGCCTGGC
TCGGCCTGGCT
CGGCCCGGCTA

CGCGACATCAC
GCGACATCACT

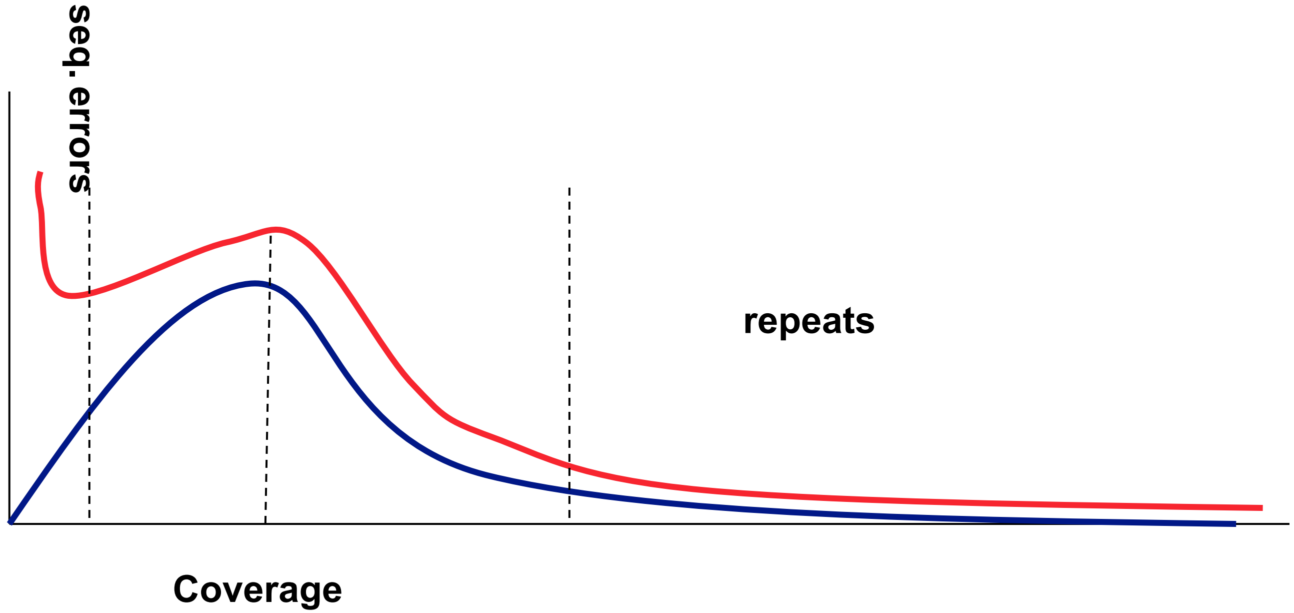
TGCGCGACATC
ATGTGAGCGAC

Overlap graph *G=(V,E)*

*V={***reads** in dataset*}*

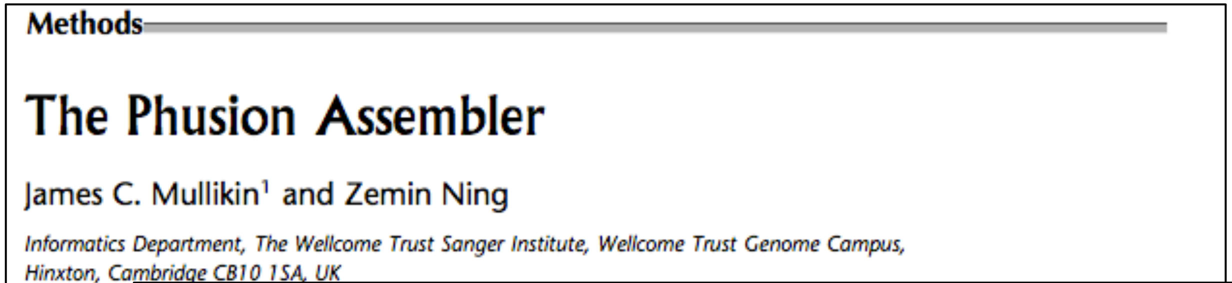
E={(r1,r2) : if r1 overlaps r2*}*

K-mer distribution



Mullikin J C , Ning Z Genome Res. 2003;13:81-90

Overlap Graphs Assembly Tools

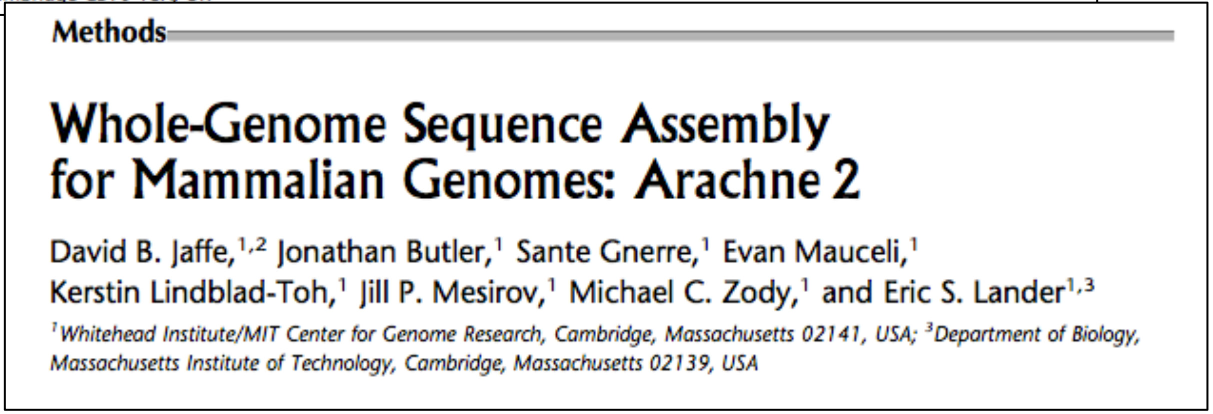


Methods

The Phusion Assembler

James C. Mullikin[1] and Zemin Ning

Informatics Department, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

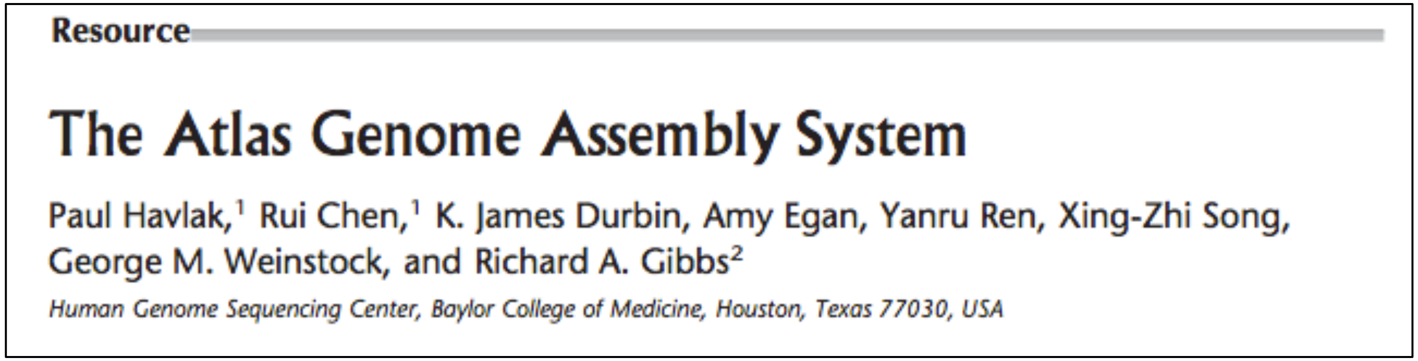


Methods

Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2

David B. Jaffe,[1,2] Jonathan Butler,[1] Sante Gnerre,[1] Evan Mauceli,[1] Kerstin Lindblad-Toh,[1] Jill P. Mesirov,[1] Michael C. Zody,[1] and Eric S. Lander[1,3]

[1] Whitehead Institute/MIT Center for Genome Research, Cambridge, Massachusetts 02141, USA; [3] Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA



Resource

The Atlas Genome Assembly System

Paul Havlak,[1] Rui Chen,[1] K. James Durbin, Amy Egan, Yanru Ren, Xing-Zhi Song, George M. Weinstock, and Richard A. Gibbs[2]

Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA

BBSRC Genome Analysis Centre
In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium

The Challenges

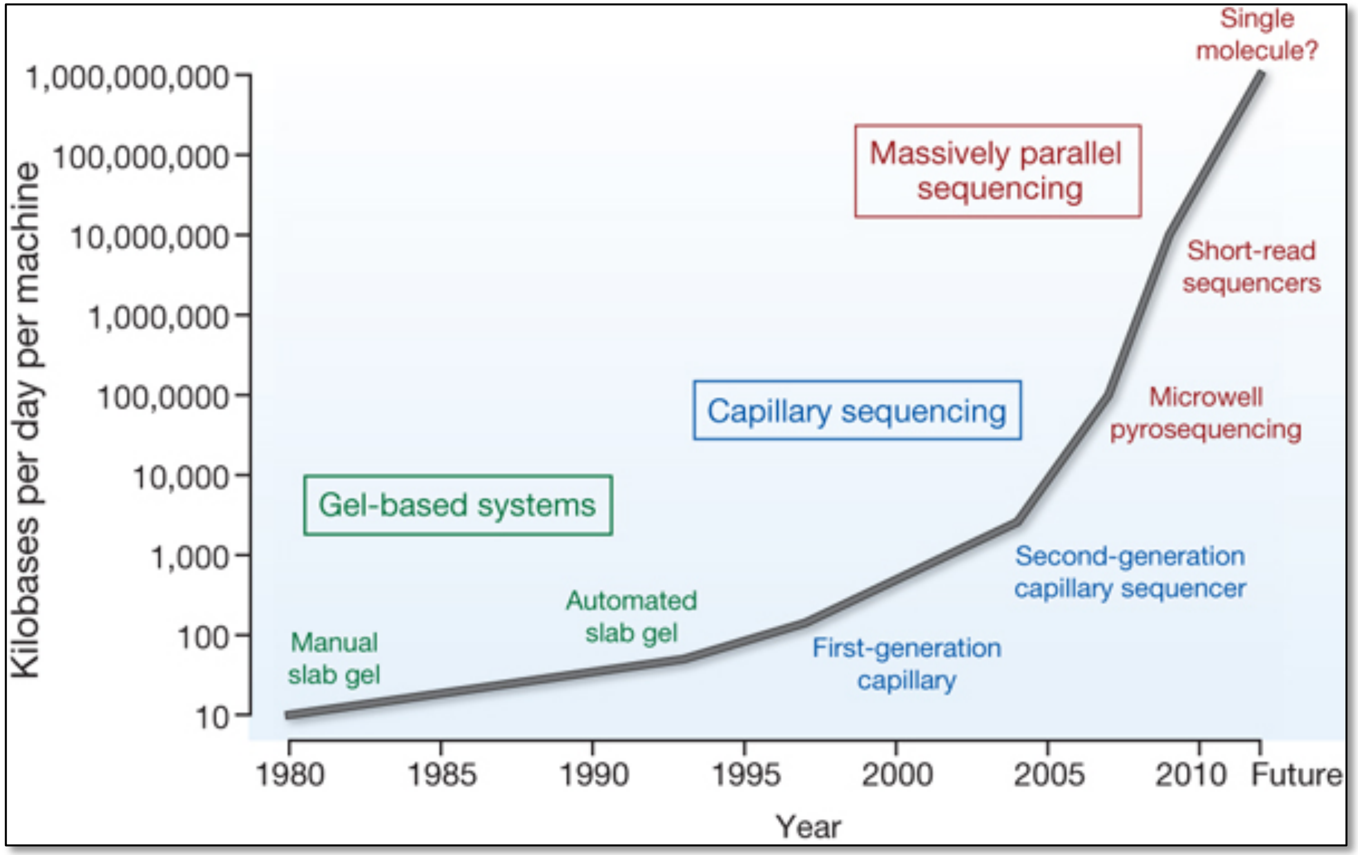
- **Genome specific**
  - base content (GC/AT)
  - repeat structure
  - homozygousity/heterozygousity

- **Technology specific**
  - number of reads
  - read length
  - sampling / sequencing bias / lack of coverage
  - memory-hungry algorithms
  - error profile
  - insert sizes

- **bioinformatics, budget, quality of samples….**

We should approach an assembly as a lab experiment.

Next Generation Technologies



Michael R. Stratton, Peter J. Campbell & P. Andrew Futreal
Nature ***458, 719-724(9 April 2009)***

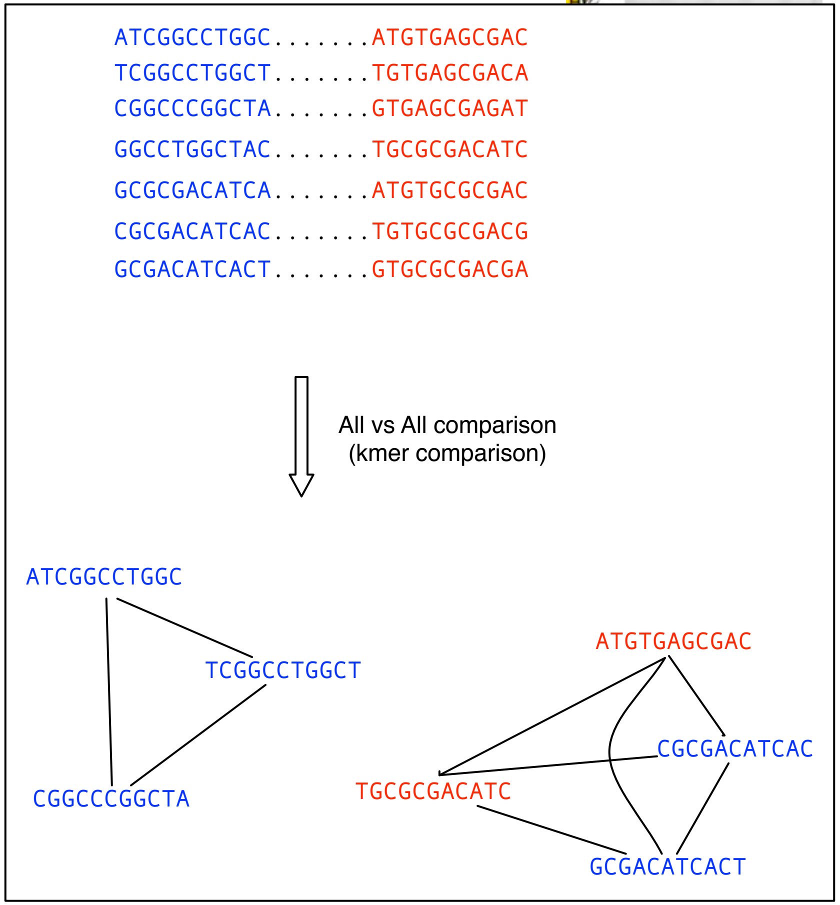
Challenge 1: far too many reads

BBSRC Genome Analysis Centre
In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium

2 x 10^9 sequence reads

Overlap graphs don't scale



ATCGGCCTGGC.......ATGTGAGCGAC
TCGGCCTGGCT.......TGTGAGCGACA
CGGCCCGGCTA.......GTGAGCGAGAT
GGCCTGGCTAC.......TGCGCGACATC
GCGCGACATCA.......ATGTGCGCGAC
CGCGACATCAC.......TGTGCGCGACG
GCGACATCACT.......GTGCGCGACGA

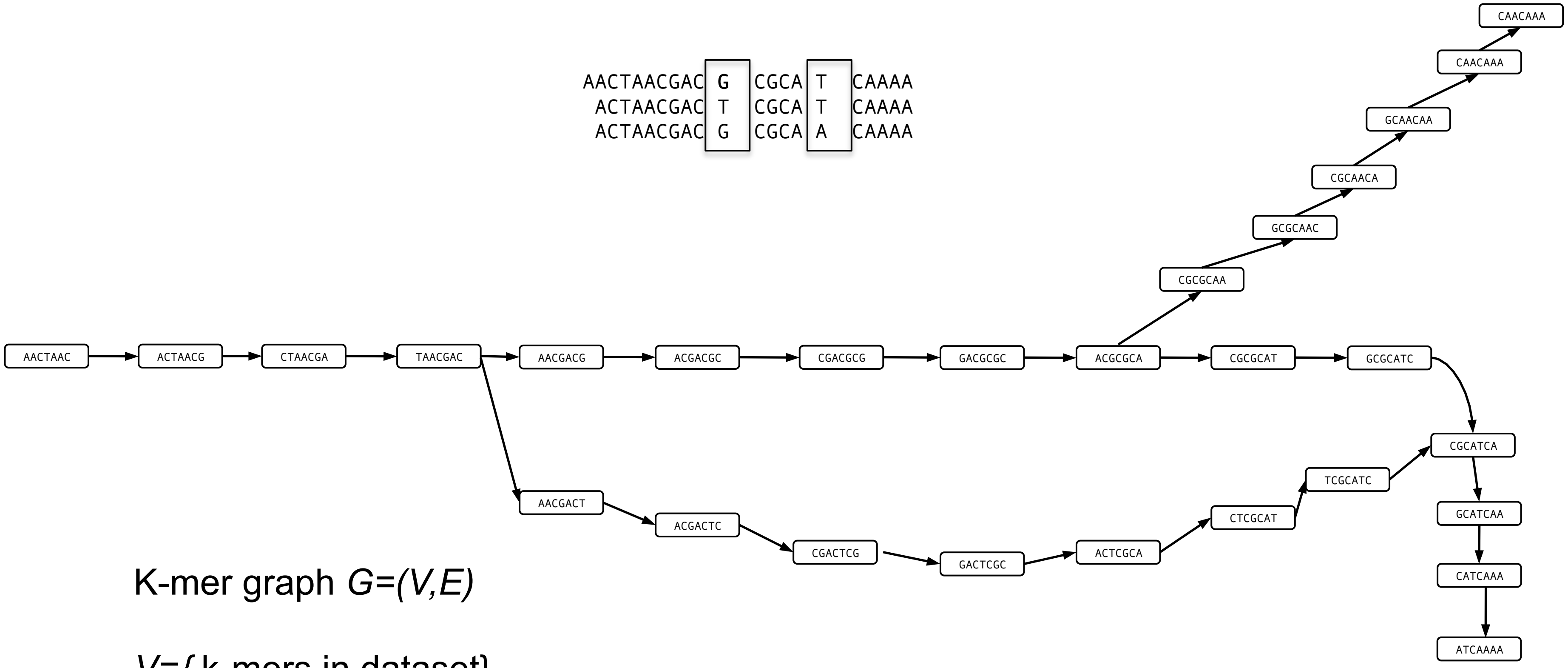
All vs All comparison
(kmer comparison)

ATCGGCCTGGC
TCGGCCTGGCT
CGGCCCGGCTA

ATGTGAGCGAC
CGCGACATCAC
TGCGCGACATC
GCGACATCACT

De Bruijn Graphs

```
AACTAACGAC G CGCA T CAAAA
 ACTAACGAC T CGCA T CAAAA
 ACTAACGAC G CGCA A CAAAA
```

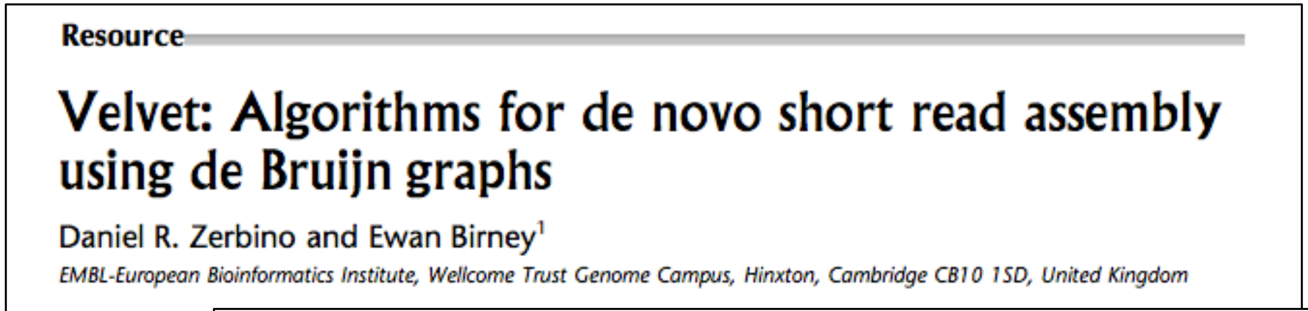


K-mer graph *G=(V,E)*

V={ k-mers in dataset}

E={(k1,k2) : if k1 overlaps k2}

Assembly tools for short-reads

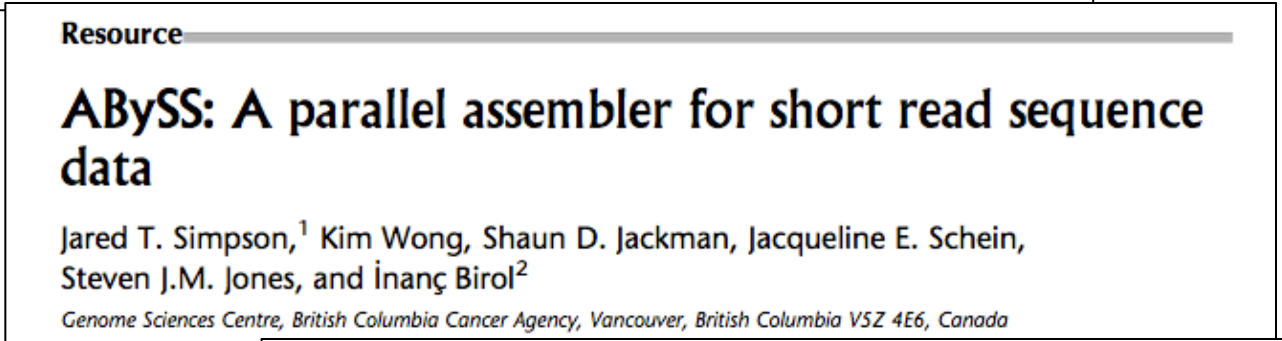


Resource

Velvet: Algorithms for de novo short read assembly using de Bruijn graphs

Daniel R. Zerbino and Ewan Birney[1]

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

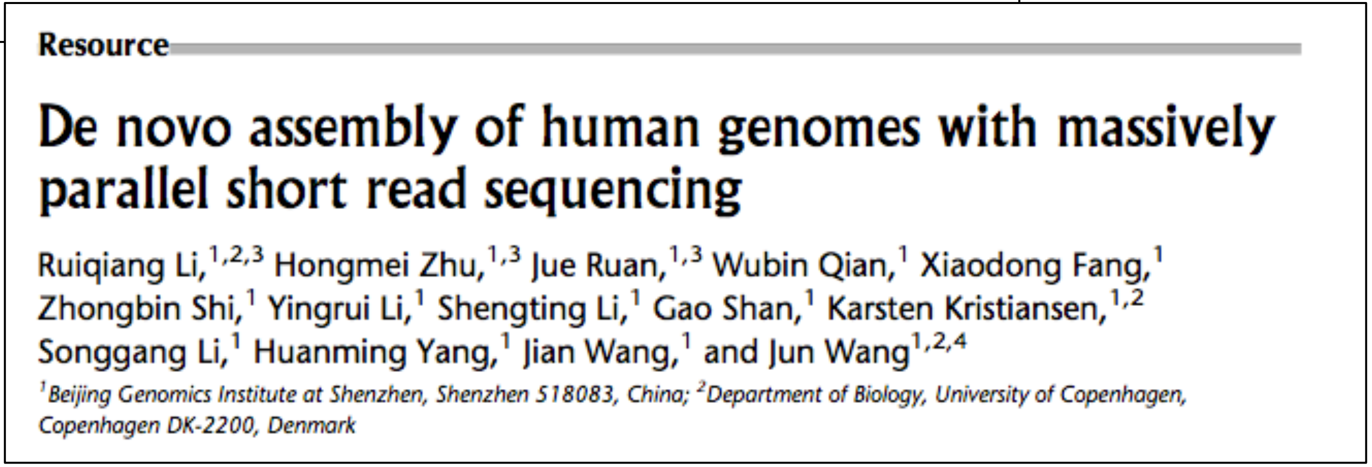


Resource

ABySS: A parallel assembler for short read sequence data

Jared T. Simpson,[1] Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, and İnanç Birol[2]

Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada



Resource

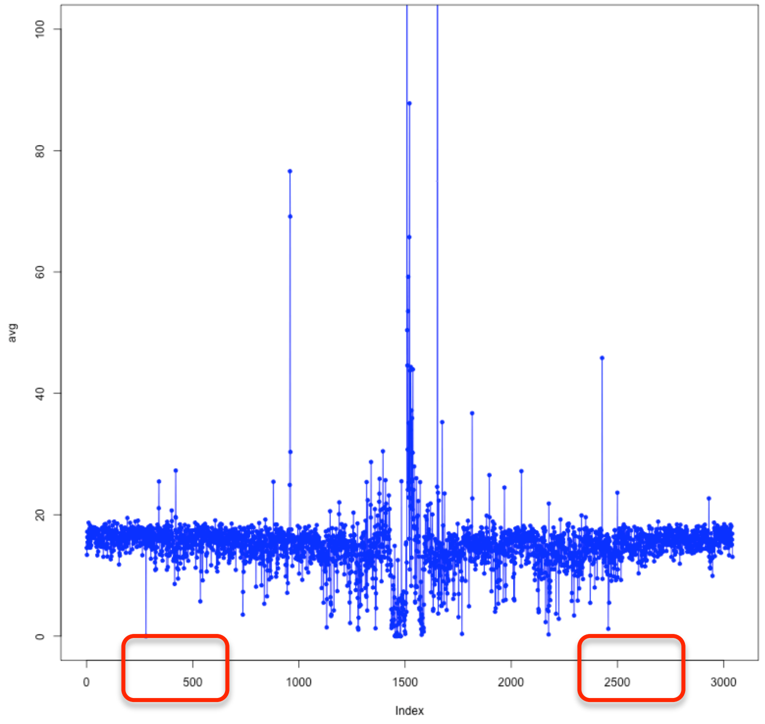
De novo assembly of human genomes with massively parallel short read sequencing

Ruiqiang Li,[1,2,3] Hongmei Zhu,[1,3] Jue Ruan,[1,3] Wubin Qian,[1] Xiaodong Fang,[1] Zhongbin Shi,[1] Yingrui Li,[1] Shengting Li,[1] Gao Shan,[1] Karsten Kristiansen,[1,2] Songgang Li,[1] Huanming Yang,[1] Jian Wang,[1] and Jun Wang[1,2,4]

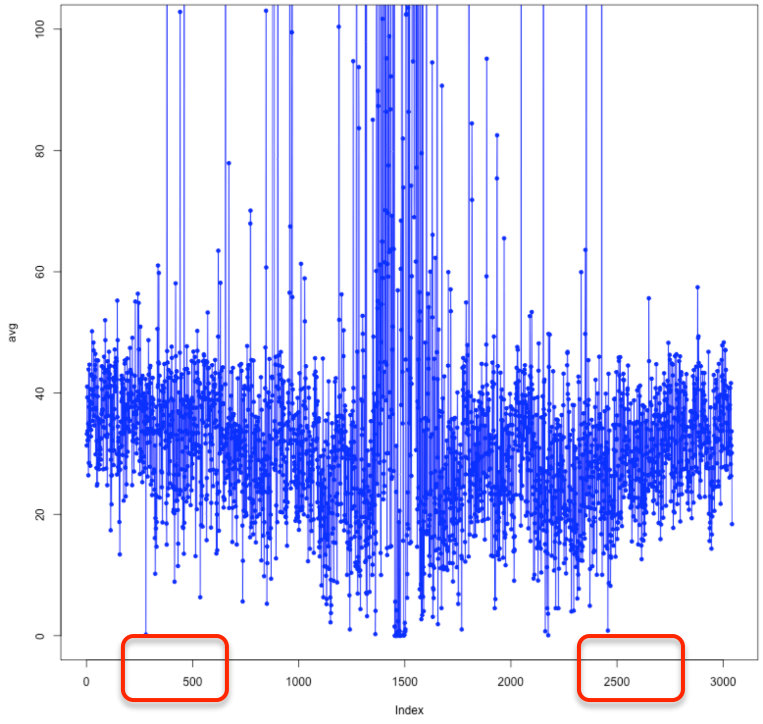
[1]Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China; [2]Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark

Challenge 2: lack of coverage

Chromosome 1 - Arabidopsis



Illumina - BWA



SOLiD – Corona Light

Nizar Drou (TGAC)

Lander-Waterman Theory

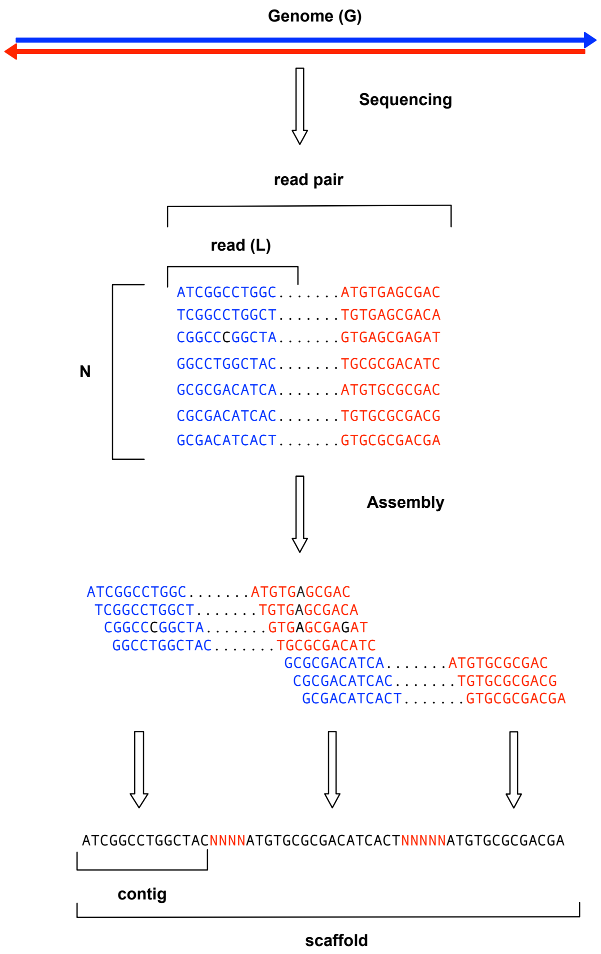
BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium

$$\text{coverage} = NL/G$$

$$P(d>k) = 1 - e^{-(NL/G)} \sum^{k} \frac{(NL/G)^k}{k!}$$



Genome (G)

Sequencing

read pair

read (L)

N

```
ATCGGCCTGGC.......ATGTGAGCGAC
TCGGCCTGGCT.......TGTGAGCGACA
CGGCCCGGCTA.......GTGAGCGAGAT
GGCCTGGCTAC.......TGCGCGACATC
GCGCGACATCA.......ATGTGCGCGAC
CGCGACATCAC.......TGTGCGCGACG
GCGACATCACT.......GTGCGCGACGA
```

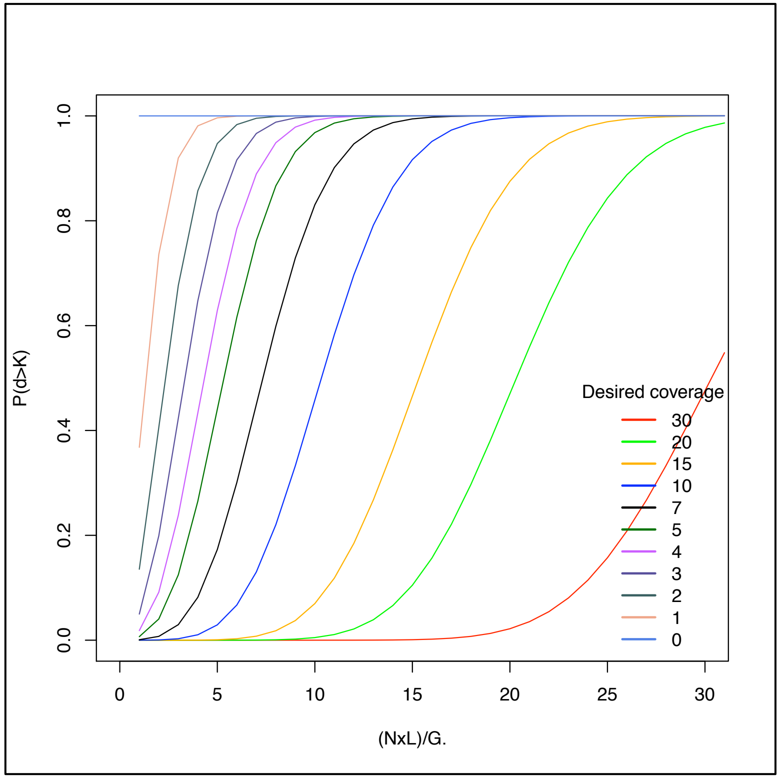
Assembly

```
ATCGGCCTGGC.......ATGTGAGCGAC
 TCGGCCTGGCT.......TGTGAGCGACA
  CGGCCCGGCTA.......GTGAGCGAGAT
   GGCCTGGCTAC.......TGCGCGACATC
                        GCGCGACATCA.......ATGTGCGCGAC
                         CGCGACATCAC.......TGTGCGCGACG
                          GCGACATCACT.......GTGCGCGACGA
```

```
ATCGGCCTGGCTACNNNNATGTGCGCGACATCACTNNNNNATGTGCGCGACGA
```

contig

scaffold

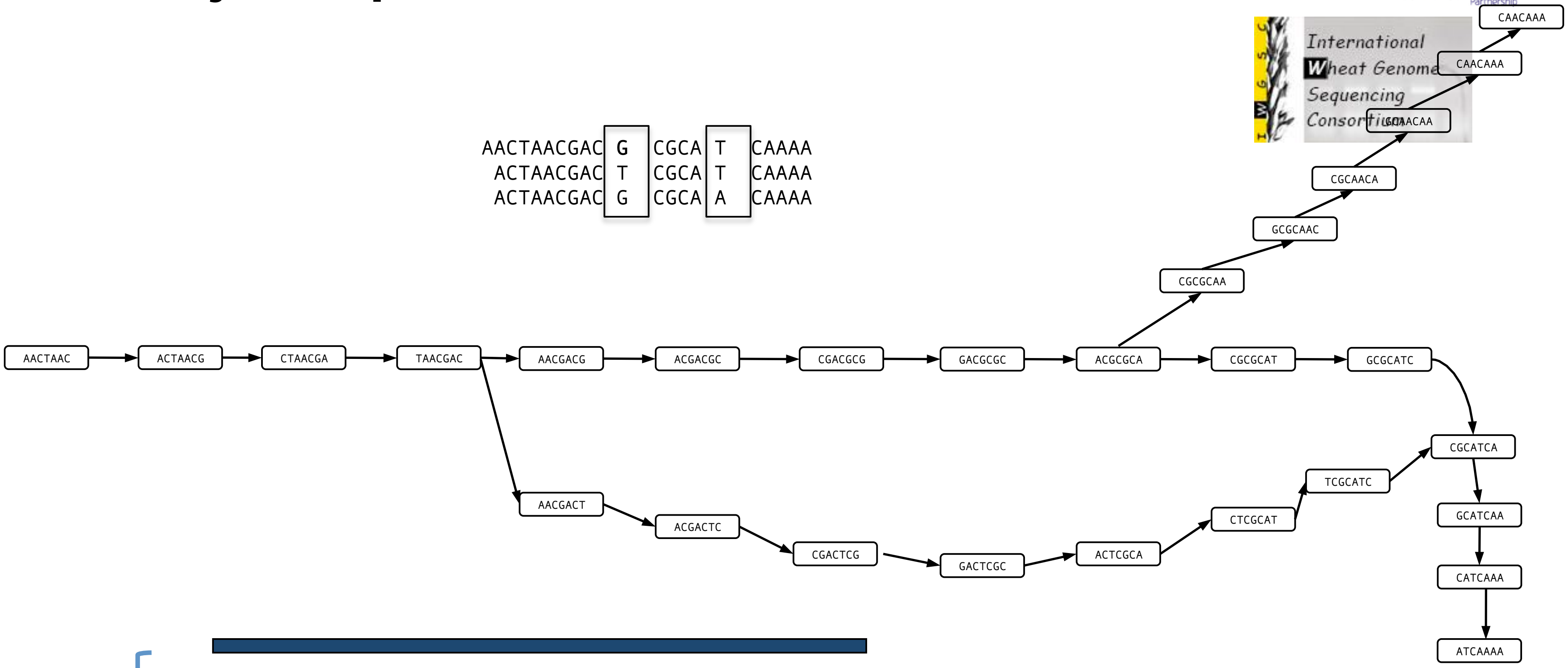


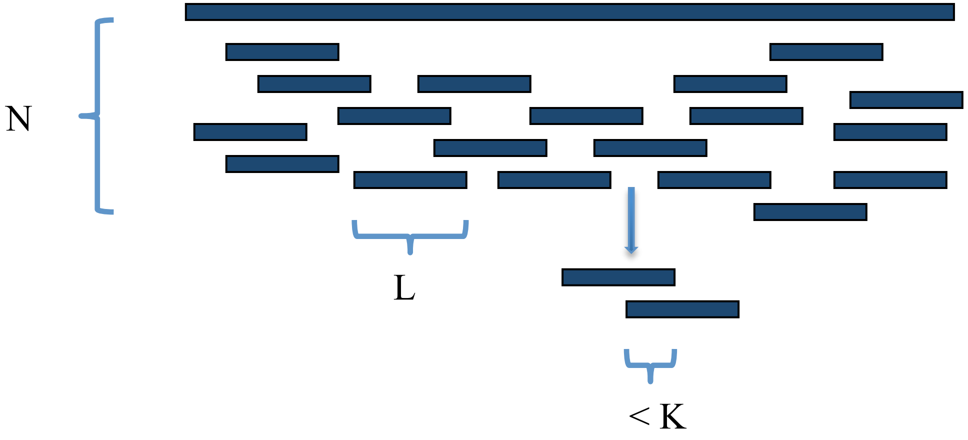
De Bruijn Graphs

BBSRC Genome Analysis Centre
In partnership with EEDA Greater Norwich Development Partnership

International Wheat Genome Sequencing Consortium

```
AACTAACGAC G CGCA T CAAAA
 ACTAACGAC T CGCA T CAAAA
 ACTAACGAC G CGCA A CAAAA
```





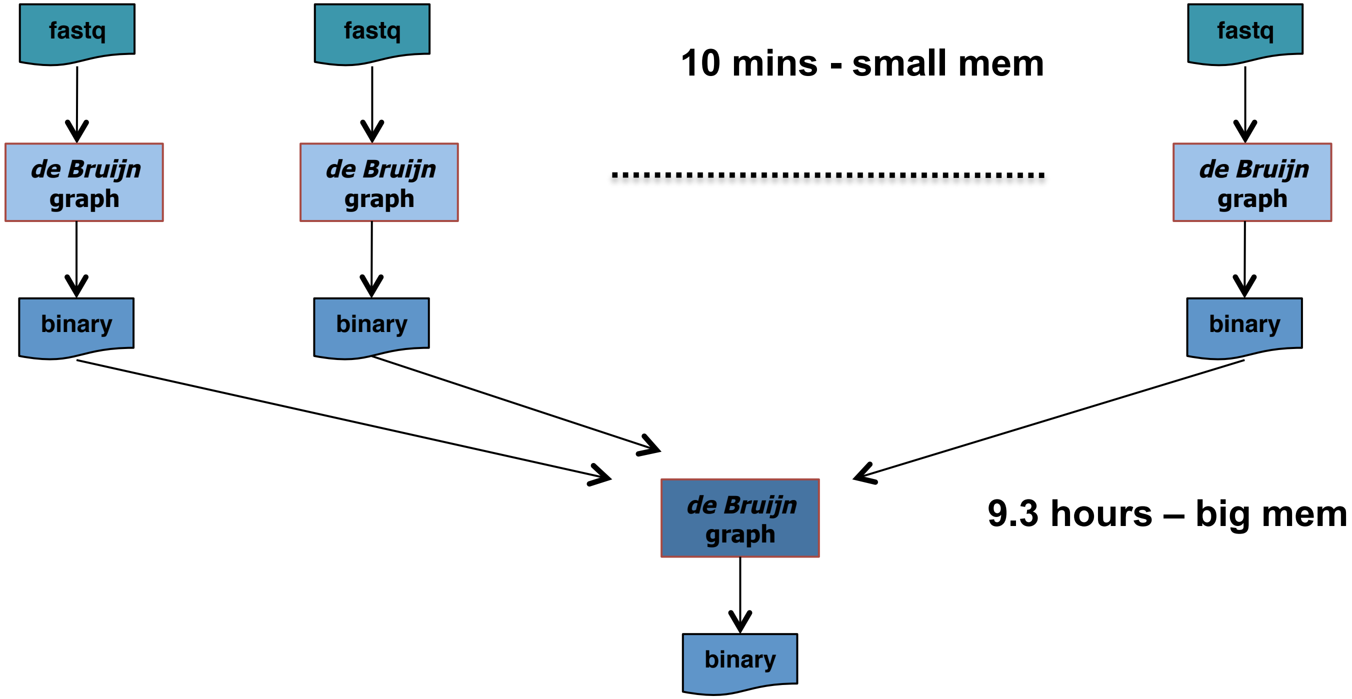
$$P(d > 0) = 1 - e^{-(N(L-K)/G)}$$

Challenge 3: memory-hungry algorithms

Cortex

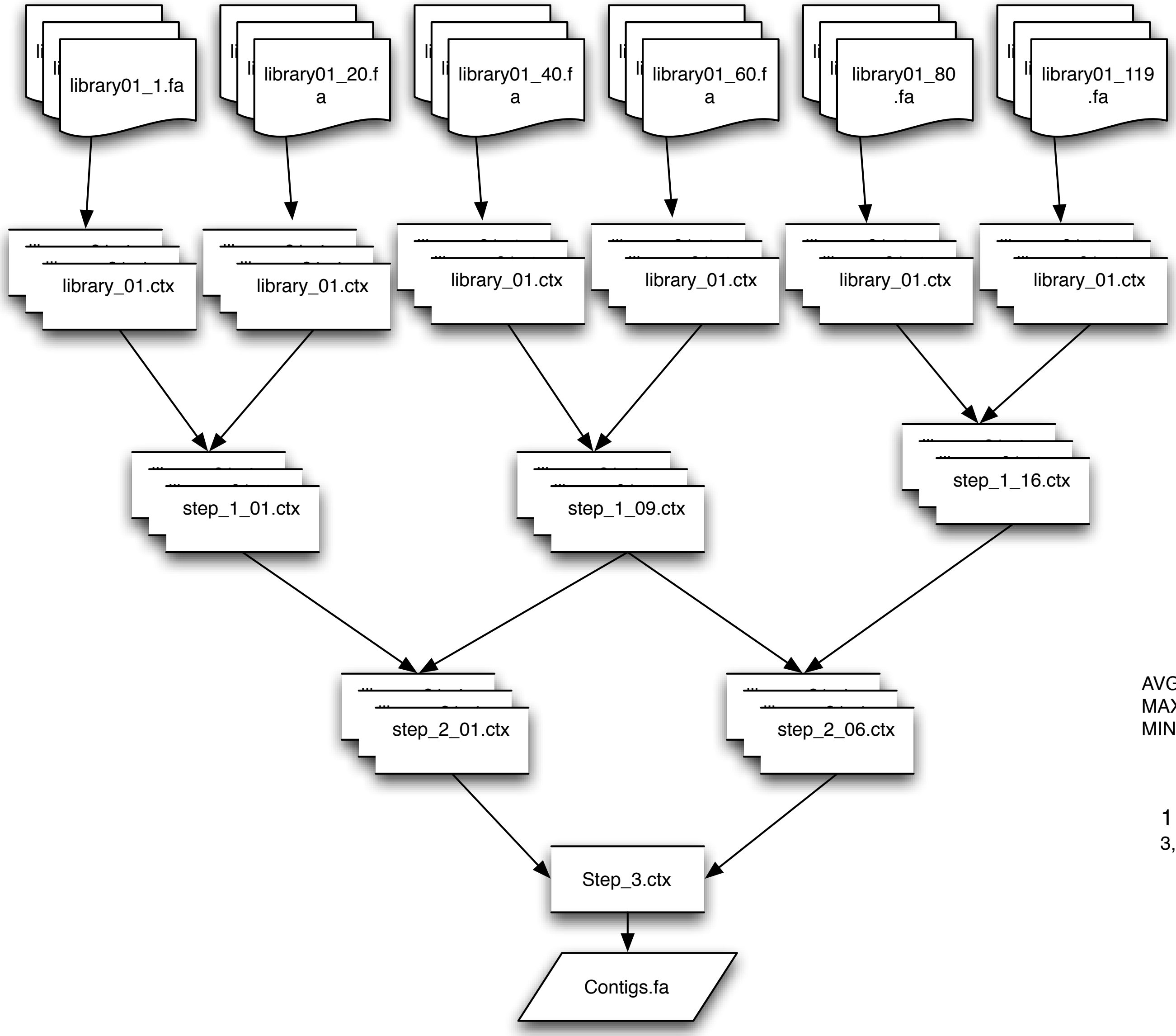
An “efficient” *de Bruijn* graph implementation (with Zamin Iqbal – Oxford)

- *de Novo* assembly (with short-reads)
- SNP/SV analysis
- Scales with number of k-mers



10 mins - small mem

9.3 hours – big mem



119 Fasta files from 454

119 Uncleaned binary files

AVG	134,767,666
MAX	211,031,386
MIN	17,223,662

16 Uncleaned binary files

AVG	568,223,963
MAX	687,341,068
MIN	353,915,818

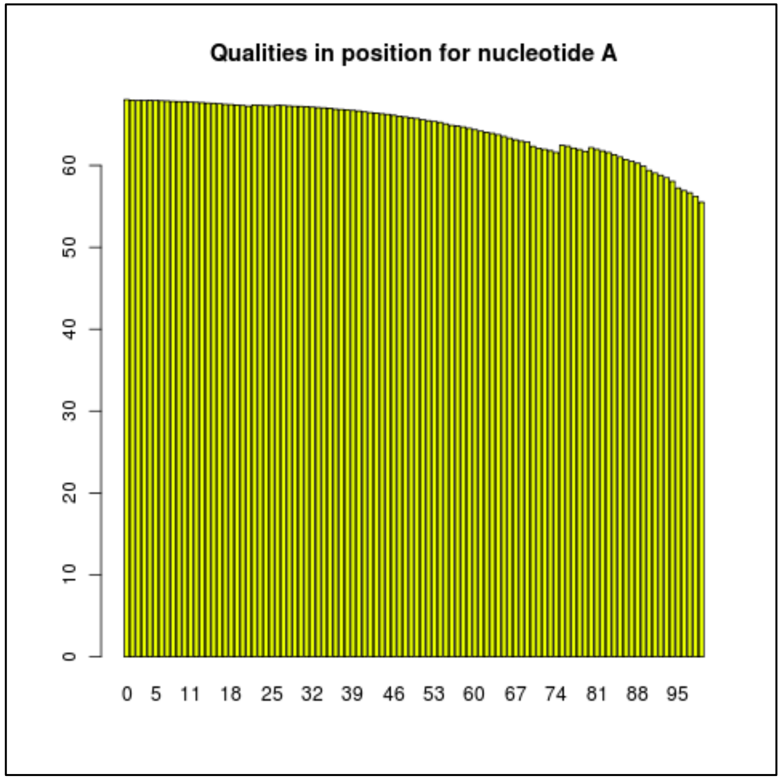
6 Cleaned binary files

AVG	1,869,897,650
MAX	2,144,187,867
MIN	1,408,682,994

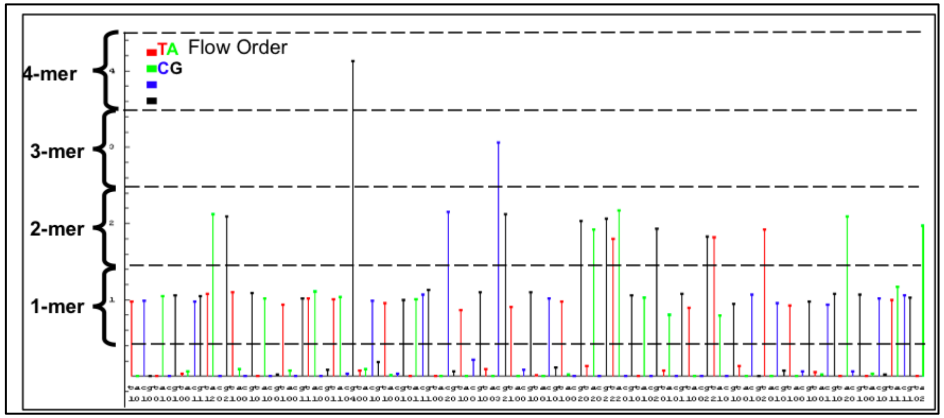
1 cleaned binary file
3,584,696,091 kmers

Ricardo Ramirez (TGAC)

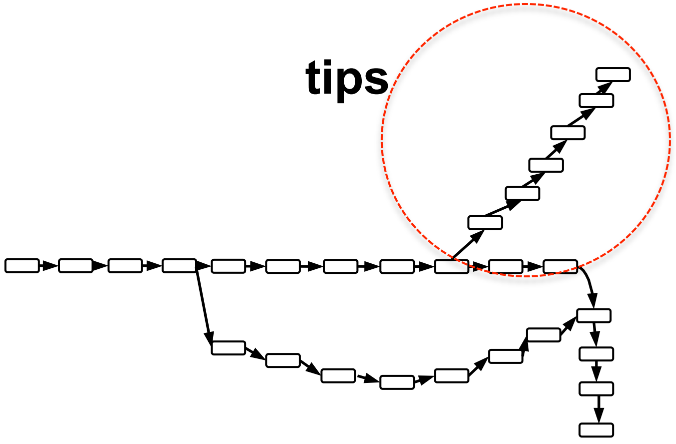
Challenge 4: error profiles



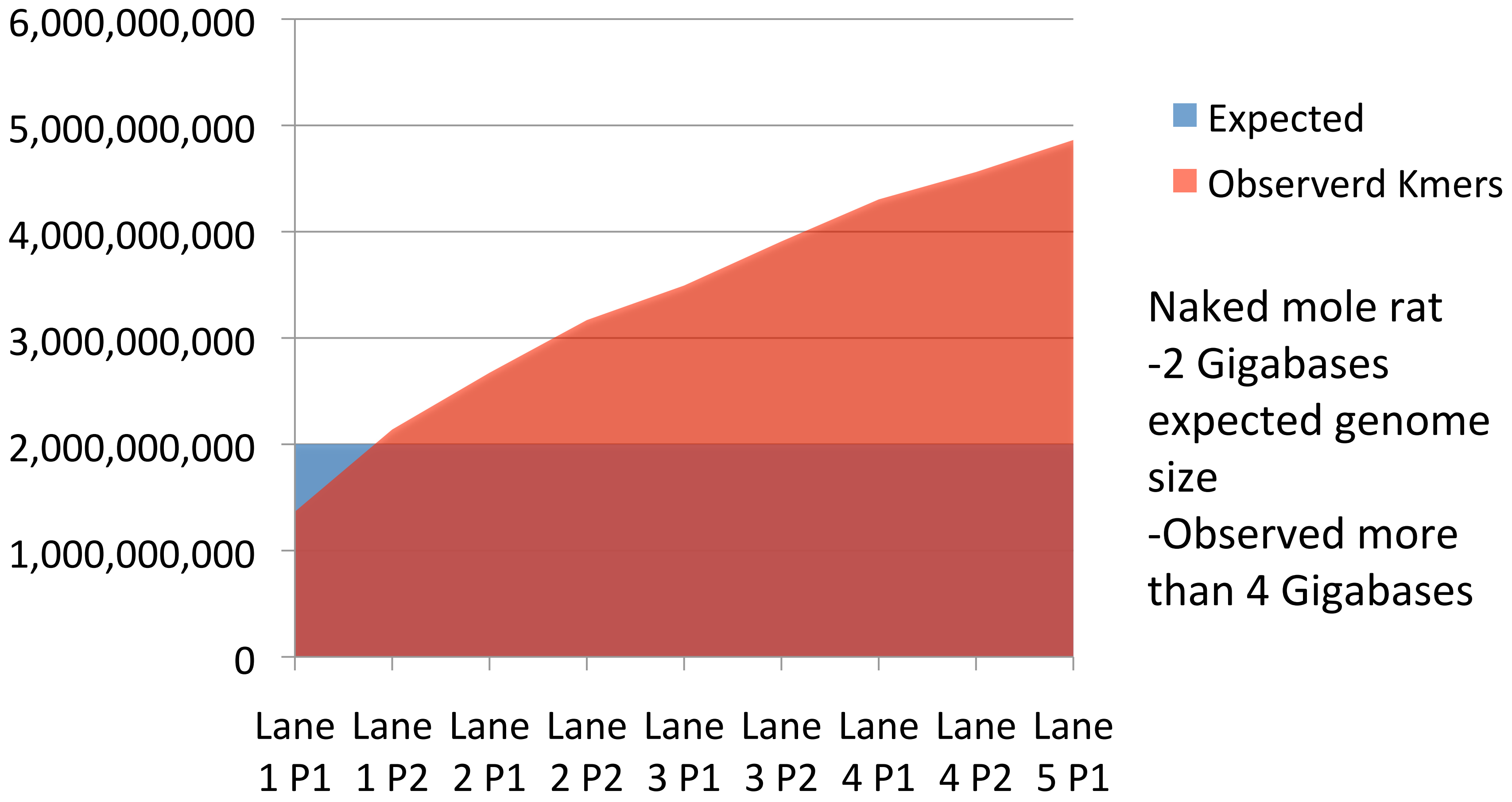
Illumina GAII



Roche 454

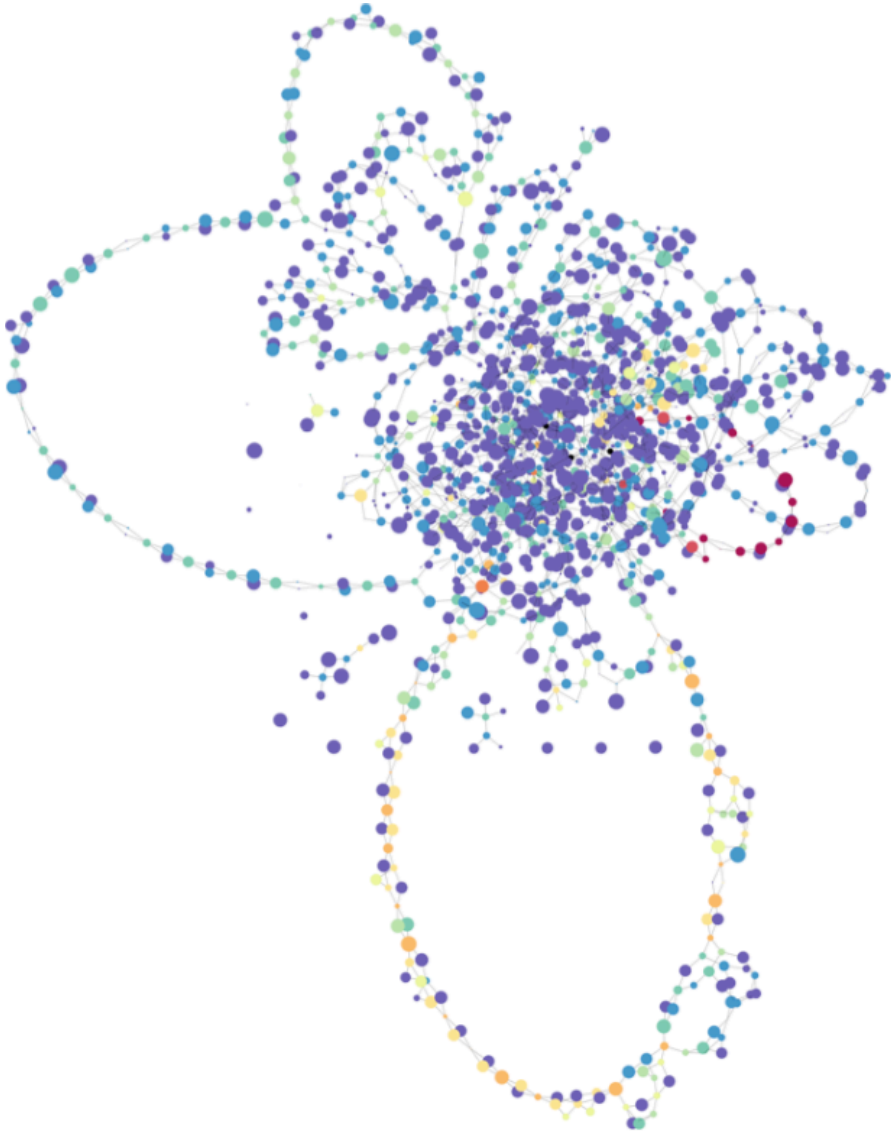


Observed K-mers vs. expected K-mers

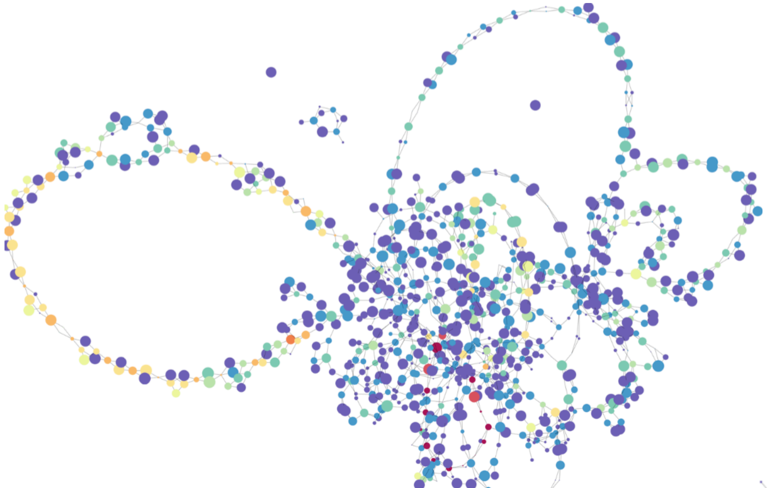


Naked mole rat
-2 Gigabases expected genome size
-Observed more than 4 Gigabases

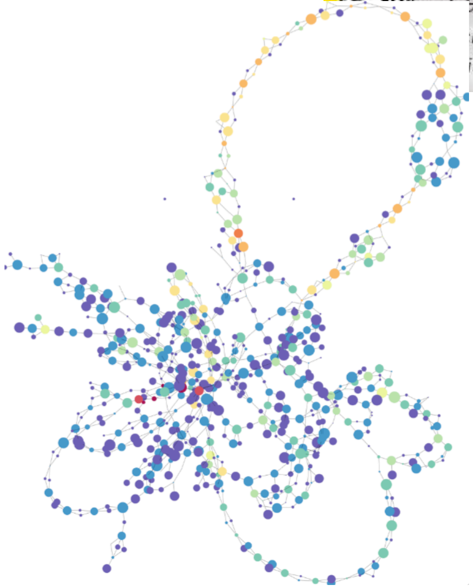
Understanding the graph



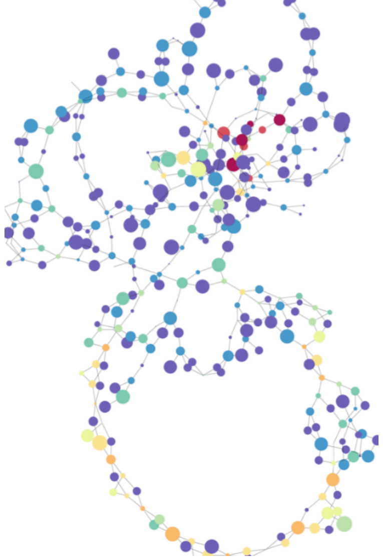
E. coli reference



K=45

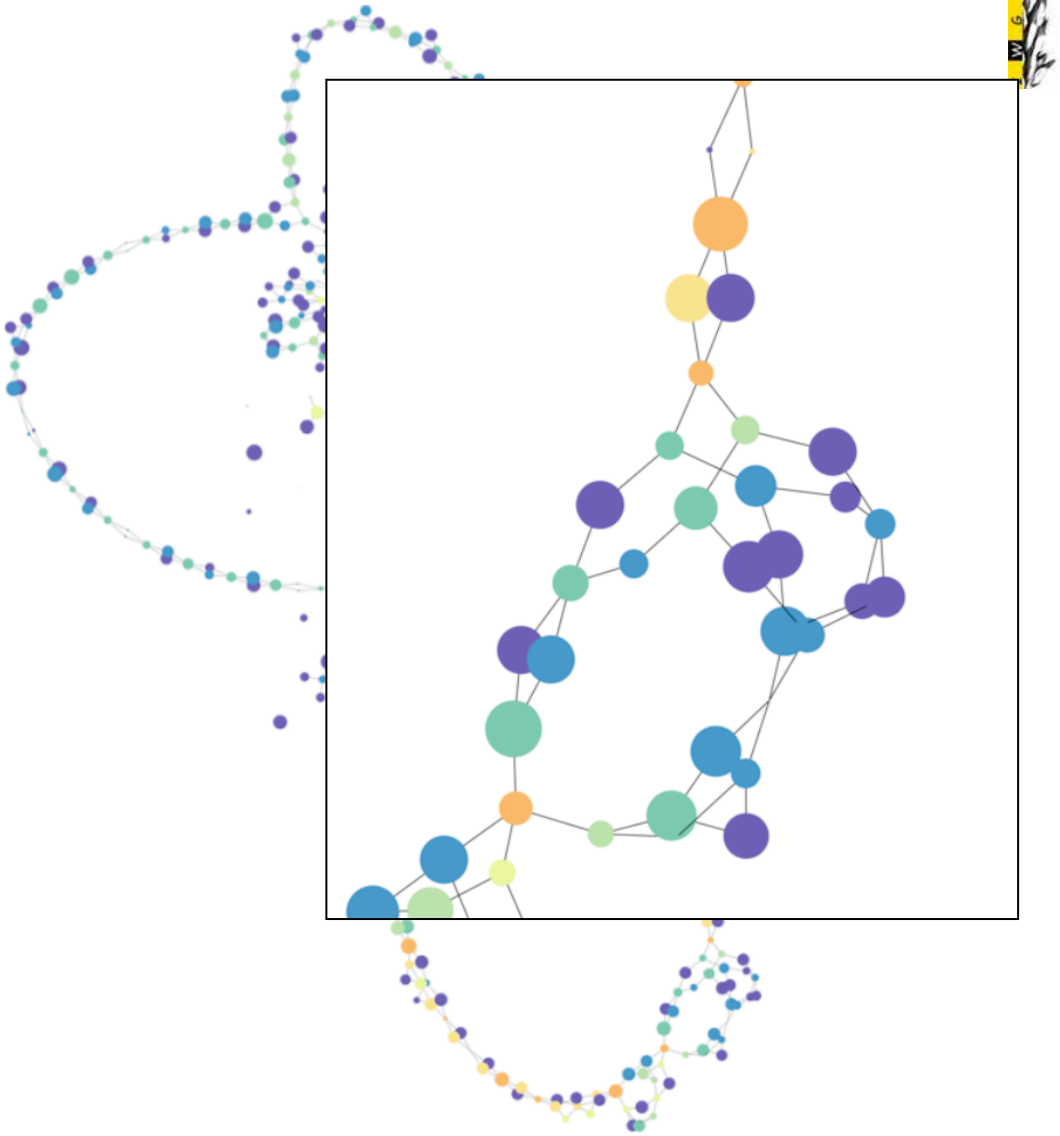


K=61



K=125

E. coli reference



Variation Analysis

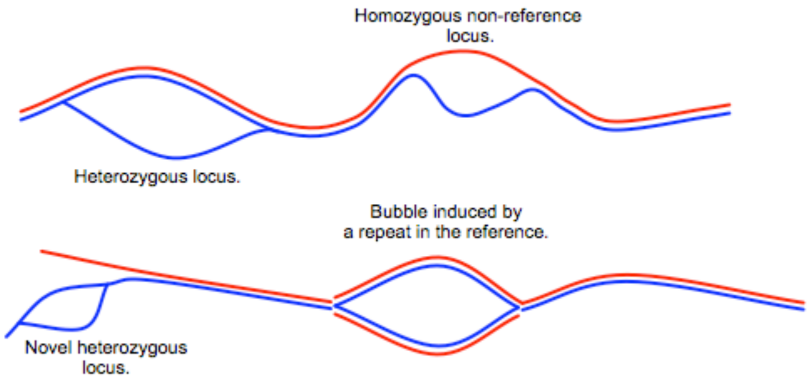

Homozygous non-reference locus.
Heterozygous locus.
Bubble induced by a repeat in the reference.
Novel heterozygous locus.

Figure 1.a

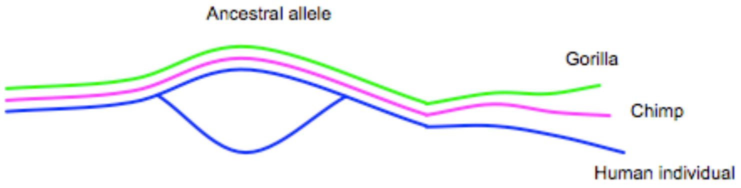


Figure 1.b

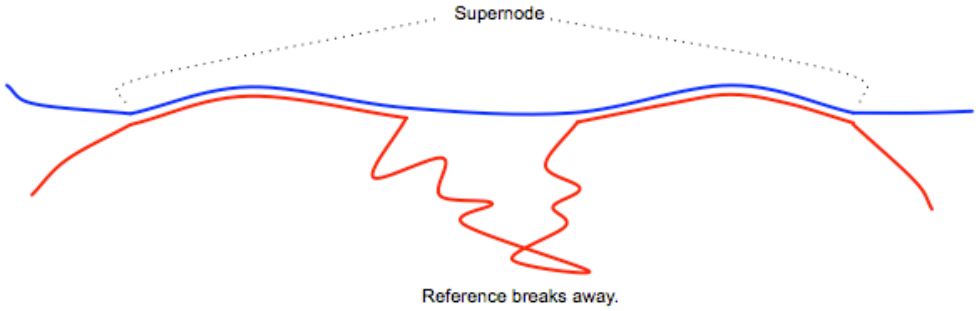


Figure 1.c

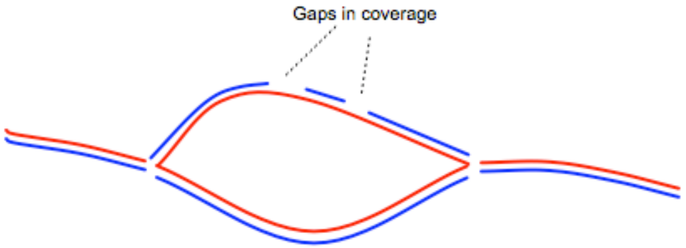


Figure 1.d

Agenda

BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

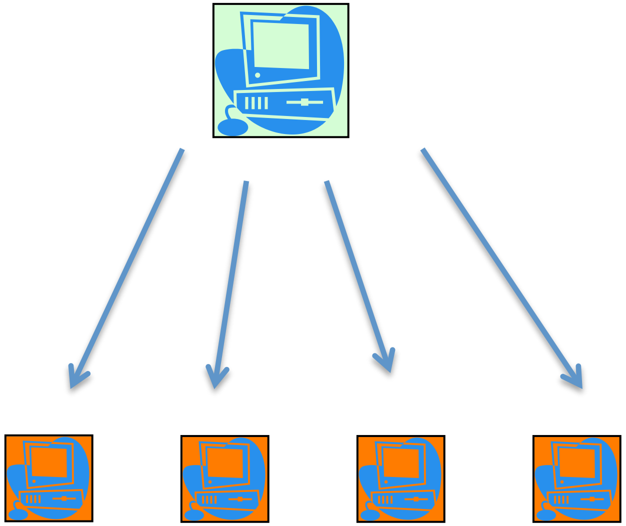
International Wheat Genome Sequencing Consortium

- Wheat Chromosome Sequencing Survey DCC
- Assemblies - theory
- Assemblies - practice

Running computing jobs in a cluster

Single Computer

Computer Cluster



Velvet

- *de novo* genomic assembler first released in 2007
- based on the de Bruijn graph approach
- developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI)
- Uses ‘Tour bus’ algorithm for tip clipping and bubble removal
- Includes the ‘Pebble’ algorithm to resolve repeats using paired end information and the ‘Rock band’ algorithm to resolve repeats when using mixed length read data, eg. reads from different platforms
- Available from http://www.ebi.ac.uk/~zerbino/velvet/

Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs .
D.R. Zerbino and E. Birney. Genome Research **18**:821-829.

Velvet (2)

First create a hashtable from a fastq file containing paired-end reads using a k-mer size of 27;

```
> velveth output_directory 27 fastq shortPaired reads.fastq
```

Generates files ‘Sequences‘ and ‘Roadmaps’ into output_directory

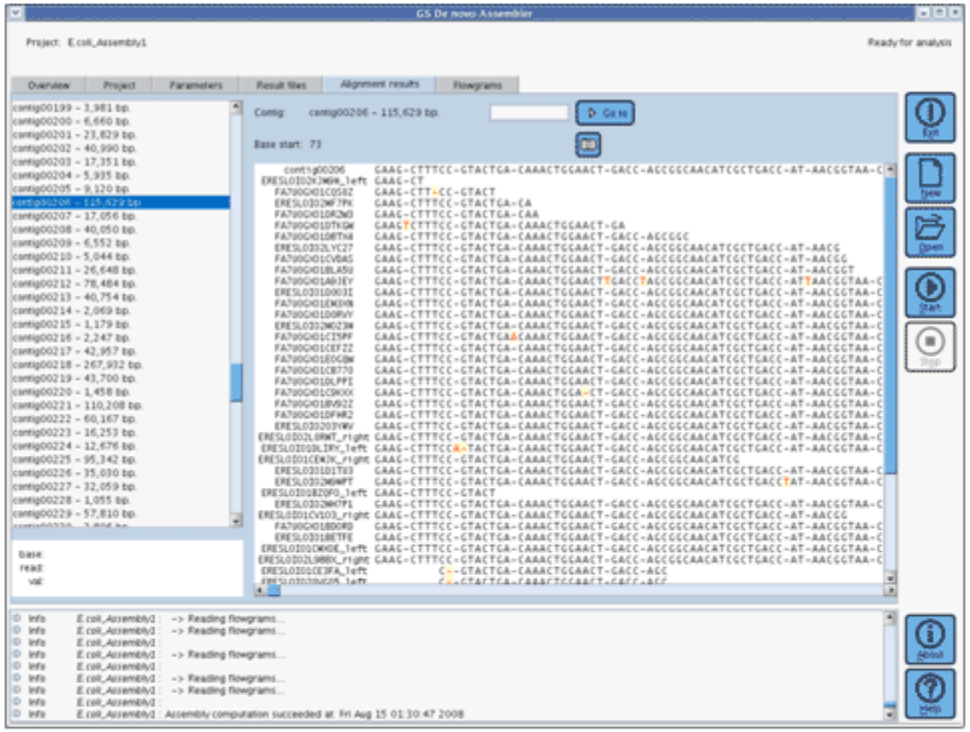
Now build and manipulate the de Bruijn graph

```
> velvetg output_directory/ -cov_cutoff 4 -min_contig_lgth 100
```

Output is contigs.fa and stats.txt

Newbler

- *de novo* assembler shipped with 454 sequencing machines
- Useful for genomes up to 3Gb in size
- Uses .sff files (the native 454 read format) or fasta with quality files
- Can be run on the command-line (runAssembly) or using the GUI interface (GS De Novo Assembler)



Project: E.coli_Assembly1

Ready for analysis

```
contig00199 - 3,981 bp.
contig00200 - 6,660 bp.
contig00201 - 23,829 bp.
contig00202 - 40,990 bp.
contig00203 - 17,351 bp.
contig00204 - 5,935 bp.
contig00205 - 9,120 bp.
contig00206 - 115,629 bp.
contig00207 - 17,056 bp.
contig00208 - 40,050 bp.
contig00209 - 6,552 bp.
contig00210 - 5,044 bp.
contig00211 - 26,648 bp.
contig00212 - 78,484 bp.
contig00213 - 40,754 bp.
contig00214 - 2,069 bp.
contig00215 - 1,179 bp.
contig00216 - 2,247 bp.
contig00217 - 42,957 bp.
contig00218 - 267,932 bp.
contig00219 - 43,700 bp.
contig00220 - 1,458 bp.
contig00221 - 110,208 bp.
contig00222 - 60,167 bp.
contig00223 - 16,253 bp.
contig00224 - 12,676 bp.
contig00225 - 95,342 bp.
contig00226 - 35,030 bp.
contig00227 - 32,059 bp.
contig00228 - 1,055 bp.
contig00229 - 57,810 bp.
```

Contig: contig00206 - 115,629 bp.

Base start: 73

```
Info    E.coli_Assembly1 :  -> Reading flowgrams...
Info    E.coli_Assembly1 :  -> Reading flowgrams...
Info    E.coli_Assembly1 :
Info    E.coli_Assembly1 :  -> Reading flowgrams...
Info    E.coli_Assembly1 :
Info    E.coli_Assembly1 :  -> Reading flowgrams...
Info    E.coli_Assembly1 :  -> Reading flowgrams...
Info    E.coli_Assembly1 :
Info    E.coli_Assembly1 : Assembly computation succeeded at: Fri Aug 15 01:30:47 2008
```

Newbler (2)

To run Newbler on a 454 read file;

```
> runAssembly -o assembly1  reads.sff
```

Results found in directory ‘assembly1’

- 454AllContigs.fna – FASTA file of contigs
- 454AllContigs.qual – Phred-based quality scores for each base in the contigs
- 454NewblerMetrics.txt – statistics from the assembly eg. number of reads and bases aligned, overlaps found, mean contig sizes

Use process_contigs.pl script to get metrics on the raw reads or on the assembly output

Also has trimming and screening options at the assembly stage to trim off primer sequences and remove vector contamination prior to assembly

Cortex

- Developed by Mario Caccamo (TGAC) and Zamin Iqbal (Oxford)
- Uses a de Bruijn graph approach incorporating efficient data structures to reduce the memory footprint
- Scales well for larger genomes (eg. wheat)
- Uses a binary format for storing intermediate graph structures allowing large genomes to be assembled in smaller sub-assemblies then recombined

Cortex (2)

Running cortex on a set of fastq files (listed in read_files) using kmer length of 27

```
cortex_con_31 --input_format fastq --input read_files
--kmer_size 27 --output_paths contigs.fa
```

Output is a set of contigs in file contigs.fa

Tip clipping and bubble removal is requested using the **--tip_clip** and **--remove_bubbles** parameters

ABySS

- a *de novo*, parallel, paired-end sequence assembler that is designed for short reads.
- Developed at Michael Smith Genome Sciences Centre (Canada)
- single-processor version is useful for assembling genomes up to 100 Mb in size.
- parallel version is implemented using MPI (message passing interface) and is capable of assembling larger genomes.

ABySS: A parallel assembler for short read sequence data. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. Genome Research, 2009-June.

ABySS (2)

Assemble reads in reads.fq using a kmer length of 25, contigs are generated in contigs.fa:

```
> ABYSS -k25 reads.fq -o contigs.fa
```

For paired-end reads:

```
> abyss-pe k=25 n=10 in='reads1.fq reads2.fq' name=ecoli
```

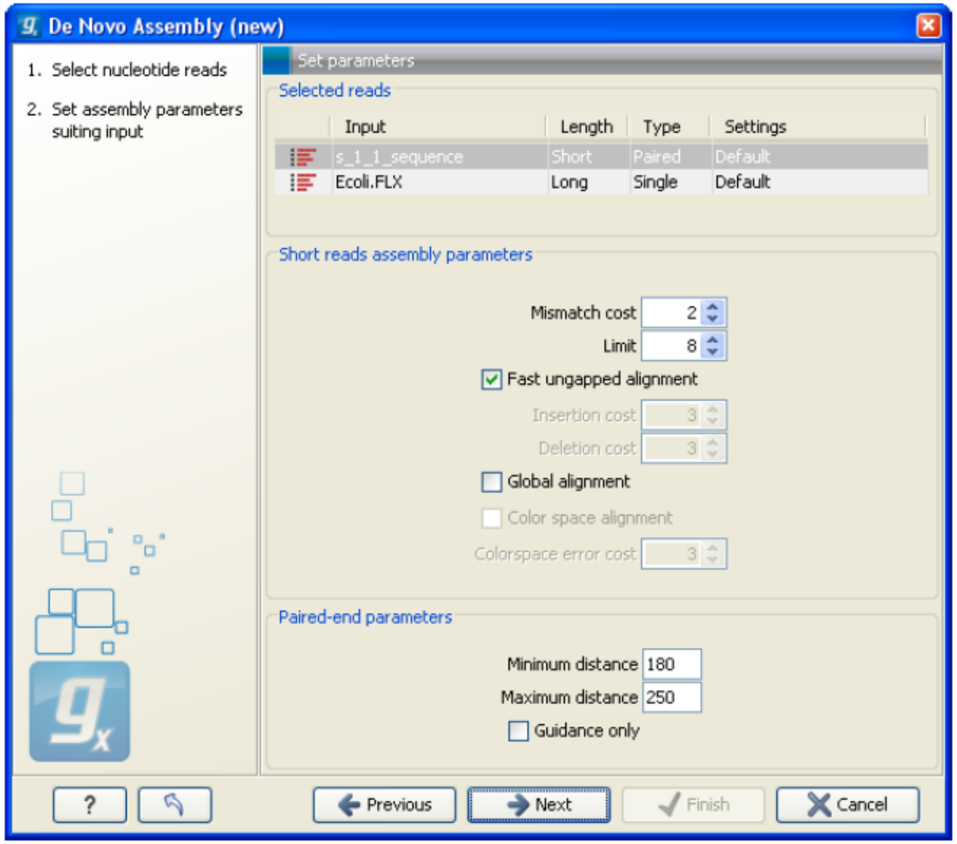
Running on a cluster using LSF:

```
> bsub -a openmpi -R "rusage[mem=75000] span[ptile=8] " -n 8
"source abyss-1.2.3 ; source openmpi-1.3.3; abyss-pe k=61 n=10
np=8 name=Name-mpi-k61 mpirun=mpirun.lsf in='reads1 ... readsN'“
```

CLC Genomics Workbench

- Commercial solution for assembly of short read data
- Developed by CLCBio, Denmark (http://www.clcbio.com)
- NOT free
- Run as a graphical interface
- Supports analysis of data from Illumina, SOLiD and 454
- de Bruijn graph based approach

CLC Genomics Workbench (2)



De Novo Assembly (new)

1. Select nucleotide reads
2. Set assembly parameters suiting input

Set parameters

Selected reads

Input	Length	Type	Settings
s_1_1_sequence	Short	Paired	Default
Ecoli.FLX	Long	Single	Default

Short reads assembly parameters

Mismatch cost 2

Limit 8

☑ Fast ungapped alignment

Insertion cost 3

Deletion cost 3

□ Global alignment

□ Color space alignment

Colorspace error cost 3

Paired-end parameters

Minimum distance 180

Maximum distance 250

□ Guidance only

- Make a table of the words seen in the reads.
- Build de Bruijn graph from the word table.
- Use the reads to resolve the repeats.
- Use the information from paired reads to resolve larger repeats.
- Output resulting contigs based on the paths.
- Contigs are available for downstream analysis through the GUI.

ALLPATHS-LG

- short read *de novo* genome assembler
- developed at the Computational Research and Development group at the Broad Institute by David Jaffe.
- Designed to assemble paired-end Illumina reads (will not assemble unpaired reads)

High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Gnerre *et al*. Proc Natl Acad Sci U S A. 2011 Jan 25;108(4):1513-8.

ALLPATHS-LG (2)

Requires reads in fastb format which are generated using a Perl script - PrepareAllPathsInputs.pl

Copy read files to a directory, eg. **/allpaths/wheat/mydata/**

Run the assembler;

```
> RunAllPathsLG PRE=allpaths DATA_SUBDIR=mydata
RUN=myrun REFERENCE_NAME=wheat TARGETS=standard K=96
```

This will create a directory under the data directory structure, eg. **/allpaths/wheat/mydata/myrun/assemblies/subdir**

The assembly files final.assembly.fasta and final.assembly.efasta are generated in subdir

Burrows-Wheeler Alignment Tool (BWA)

- Aligns relatively short sequences (queries) to a sequence database, eg. a reference genome
- Based on Burrows-Wheeler Transform (BWT).
- Developed by Heng Li at the Sanger Institute (who also developed MAQ)
- The algorithm is designed for short queries up to ~200bp with low error rate (<3%).
- Performs gapped global alignment w.r.t. queries and supports paired-end reads
- One of the fastest short read alignment algorithms to date.
- Supports colorspace alignment (SOLiD reads)
- Supports the Sequence Alignment/Map (SAM) format

BWA (2)

Index the database (fasta file)

```
> bwa index -a bwtsw database.fasta
```

Find the suffix array coordinates of the input reads

```
> bwa aln database.fasta short_read.fastq > aln_sa.sai
```

Generate alignments in SAM format (single reads)

```
> bwa samse database.fasta aln_sa.sai short_read.fastq > aln.sam
```

Generate alinments in SAM format (paired reads reads)

```
bwa sampe database.fasta aln_sa1.sai aln_sa2.sai read1.fq
read2.fq > aln.sam
```

Use SAMTools or BioPerl scripts to analyse alignment files

Bowtie

- An ultrafast, memory-efficient short read aligner
- Developed by Steven Salzberg at the University of Maryland Centre for Bioinformatics and Computational Biology
- Indexes the genome with a Burrows-Wheeler index to keep its memory footprint small
- Supports colorspace alignment (SOLiD reads)
- Supports the Sequence Alignment/Map (SAM) format

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.

Bowtie (2)

Index the reference

```
> bowtie-build –f reference.fasta e_coli
```

Align your paired-end reads and output alignments in SAM format

```
> bowtie -q -s e_coli -1 reads1.fastq -2 reads2.fastq alignments.sam
```

BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

THE END