Web-based Tools for the Integration of Genomic Data

Mario Caccamo (TGAC)

The Genome Analysis Centre

A new facility to provide critical mass and excellence in genomics specialised in **animal, microbial** and **plant** research:

- high throughput sequencing
- new technology platforms
- bioinformatics
- impact through innovation and enterprise

www.tgac.bbsrc.ac.uk

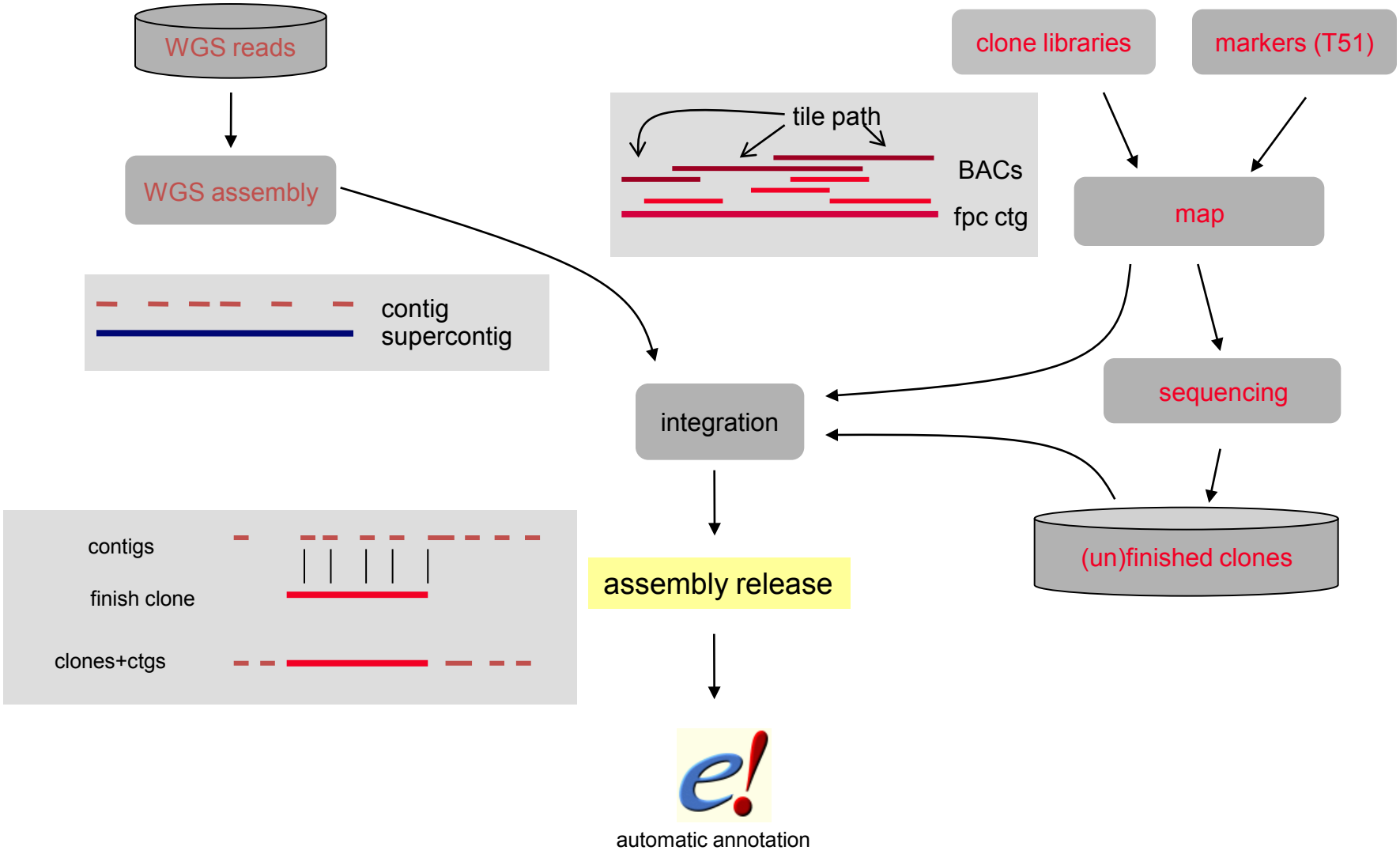
BBSRC Genome Analysis Centre

In partnership with EEDA Greater Norwich Development Partnership

Zebrafish Genome Project

whole genome shotgun sequencing

clone mapping and sequencing



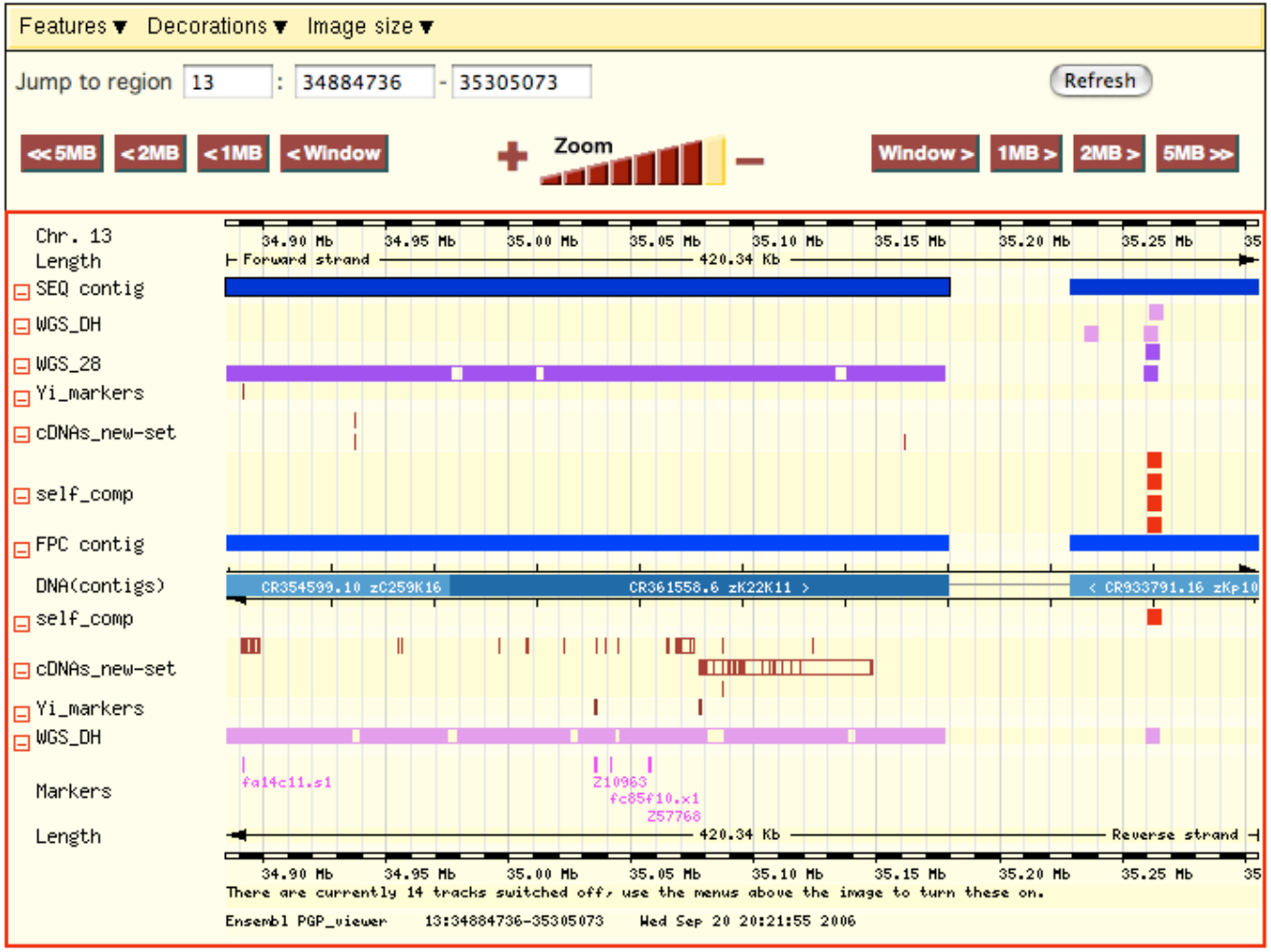
Sequence Analysis Tools

Design and implementation of tools to analyse, integrate and visualise the available resources

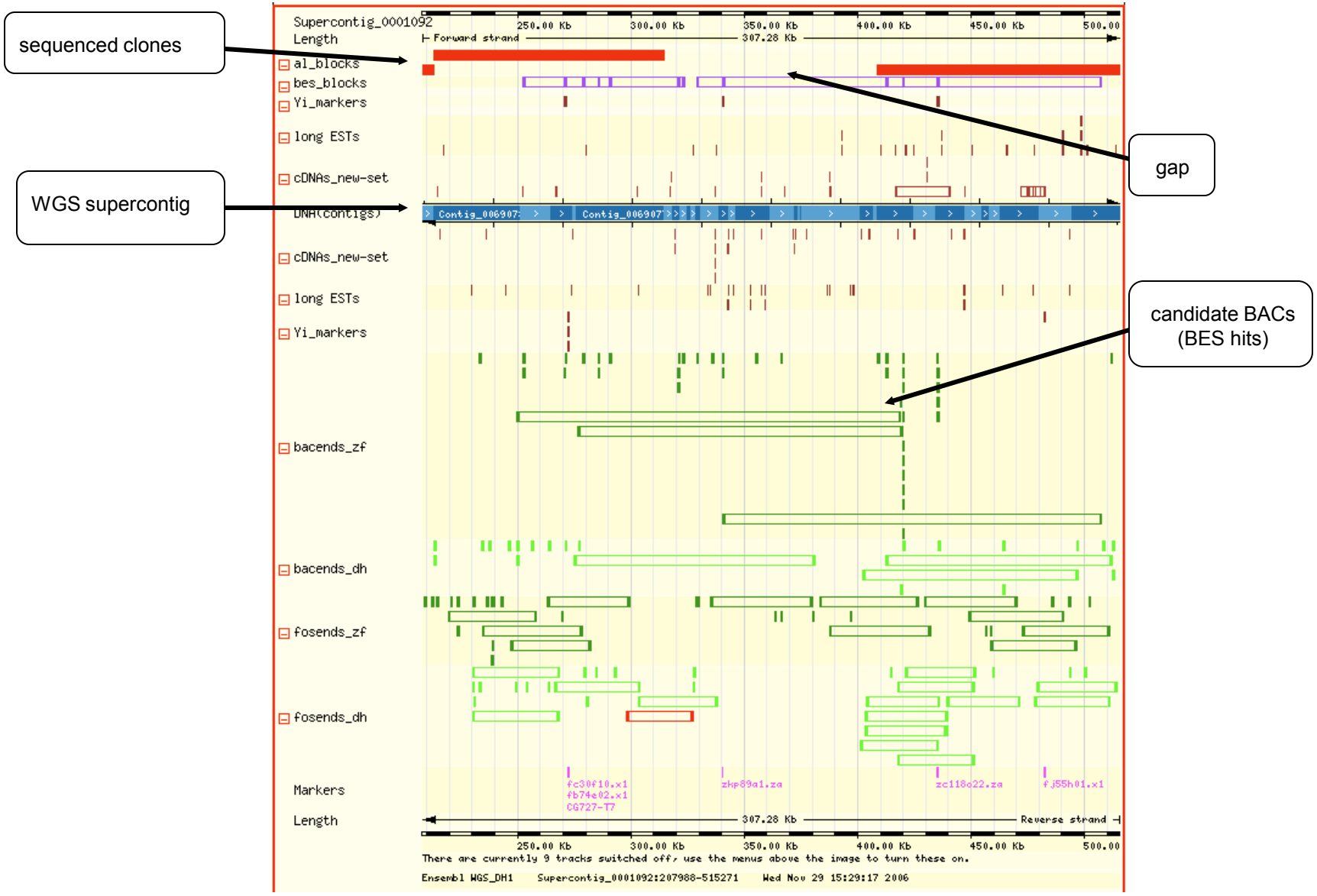
- PGP-viewer and WGS browsers
- “Punchlists”
- Data Analysis

  eg. analysis of gaps, missing cDNAs, integration of WGS data

PGP Viewer


Features ▼ Decorations ▼ Image size ▼
Jump to region 13 : 34884736 - 35305073
Refresh
<<5MB <2MB <1MB <Window
Zoom
Window > 1MB > 2MB > 5MB >>
Chr. 13
Length
34.90 Mb 34.95 Mb 35.00 Mb 35.05 Mb 35.10 Mb 35.15 Mb 35.20 Mb 35.25 Mb
Forward strand
420.34 Kb
SEQ contig
WGS_DH
WGS_28
Yi_markers
cDNAs_new-set
self_comp
FPC contig
DNA(contigs)
CR354509.10 zC259K16
CR361558.6 zK22K11 >
< CR933791.16 zKp10
self_comp
cDNAs_new-set
Yi_markers
WGS_DH
Markers
fa14c11.s1
Z10963
fc85f10.x1
Z57768
Length
420.34 Kb
Reverse strand
There are currently 14 tracks switched off; use the menus above the image to turn these on.
Ensembl PGP_viewer 13:34884736-35305073 Wed Sep 20 20:21:55 2006

Bridging Gaps with WGS Sequences



Mapped 1 time

Mapped multiple times

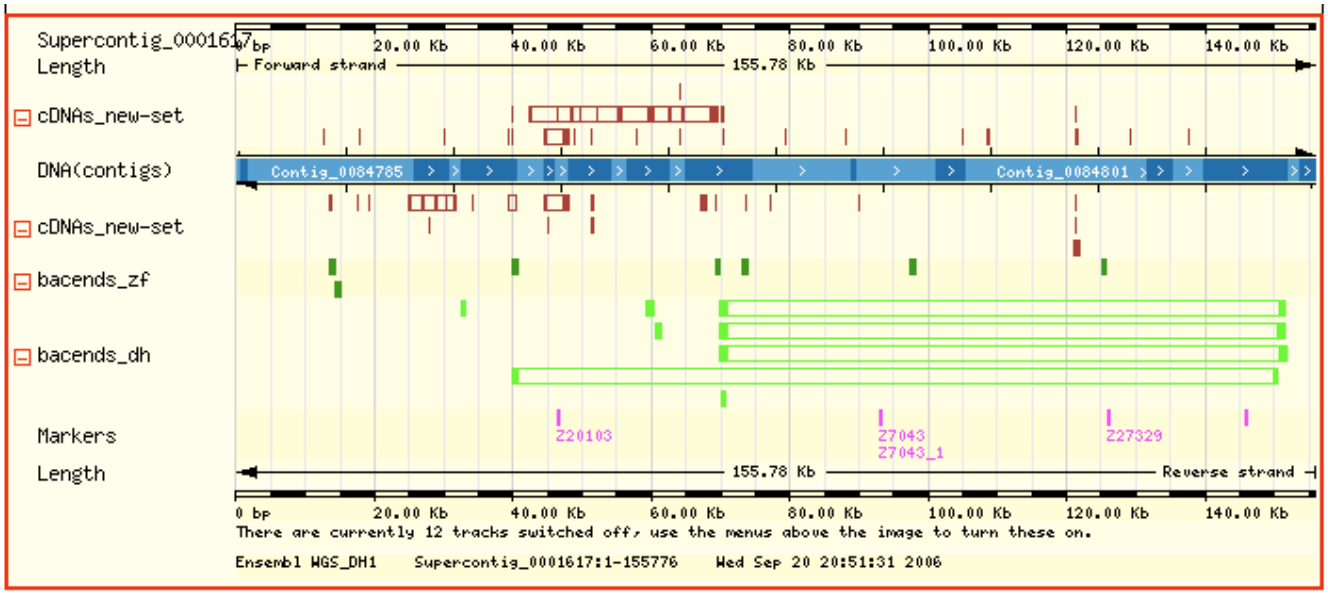
Wrong direction (<<, <>, >>)

Wrong distance

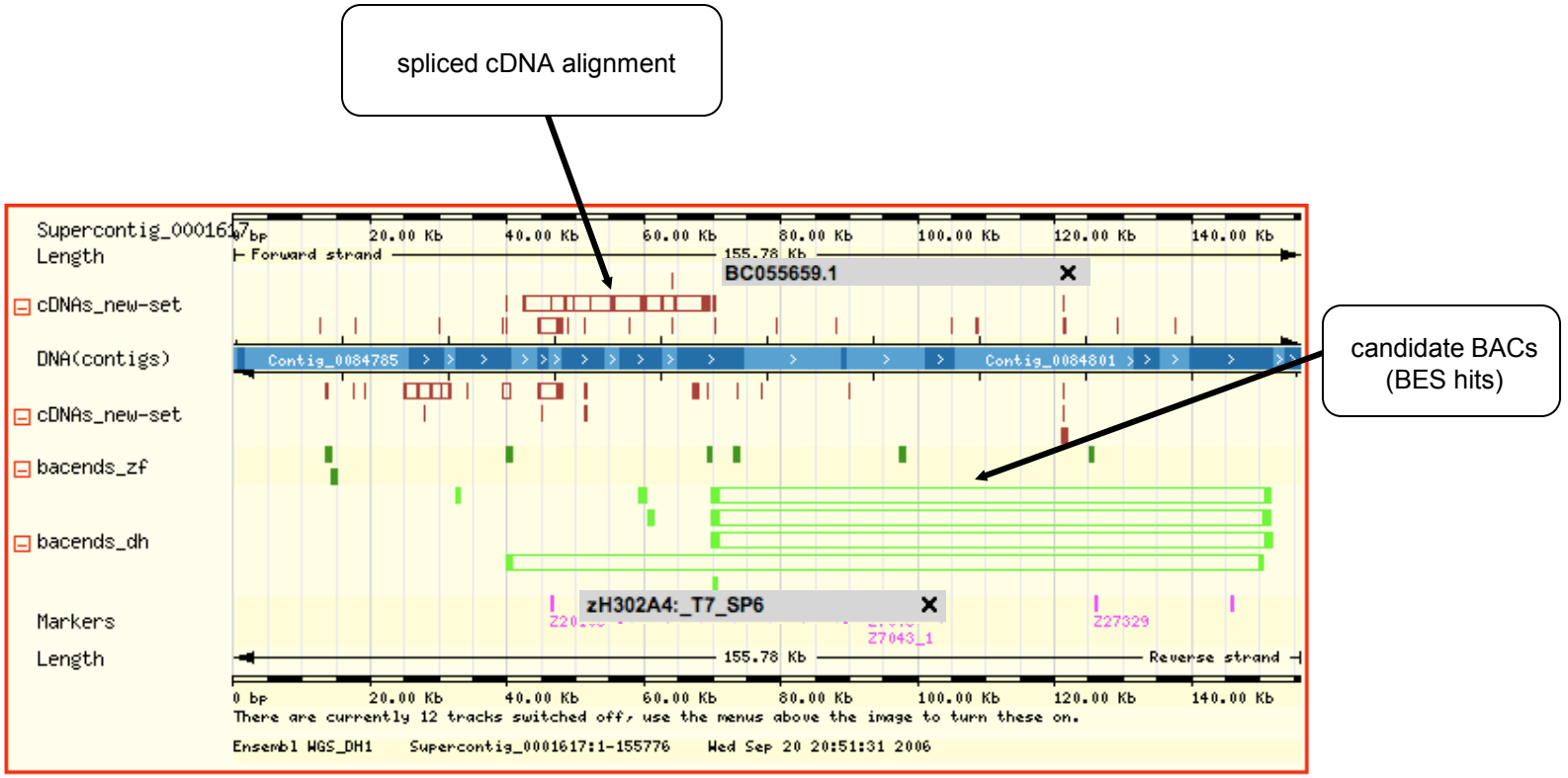
Spanner parter in the vicinity

Missing Markers/cDNAs

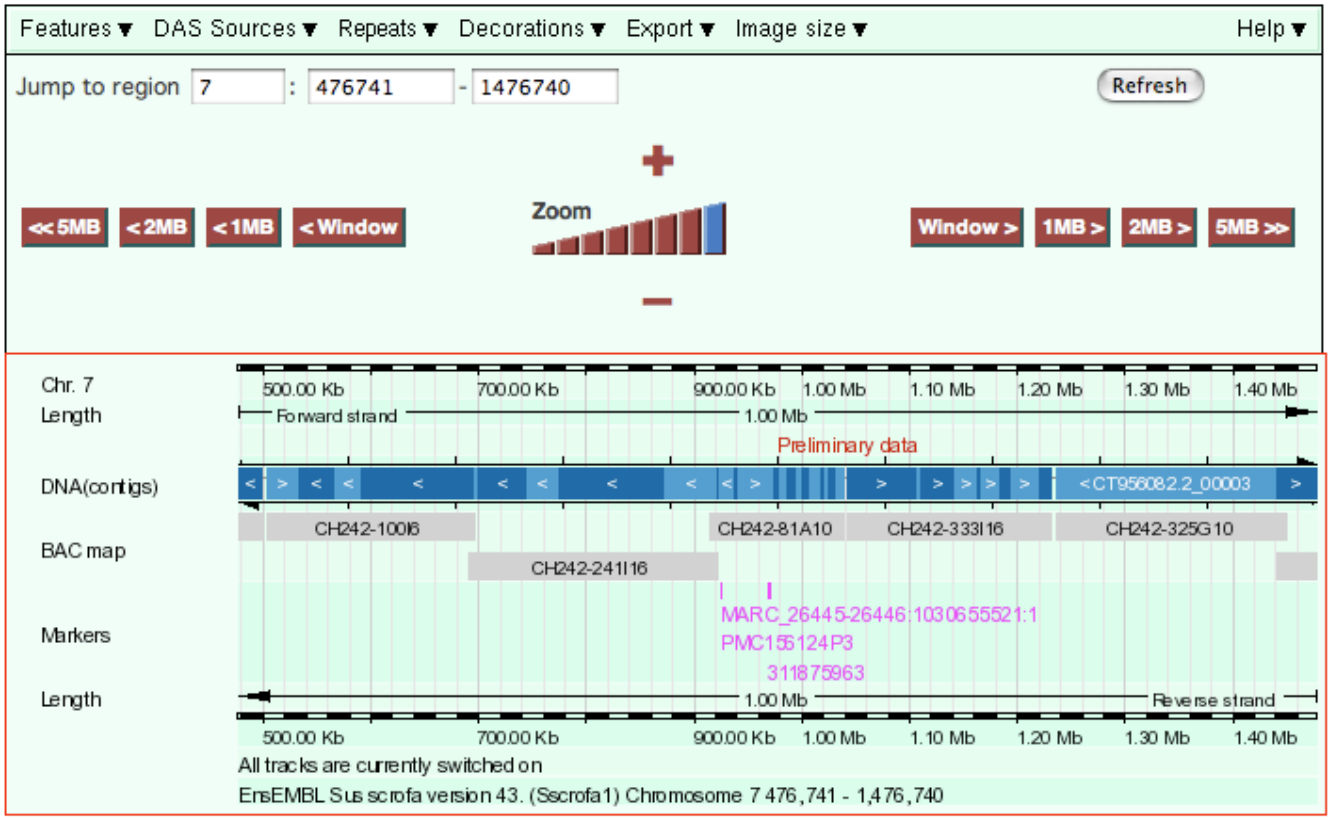
Region not covered by the physical map



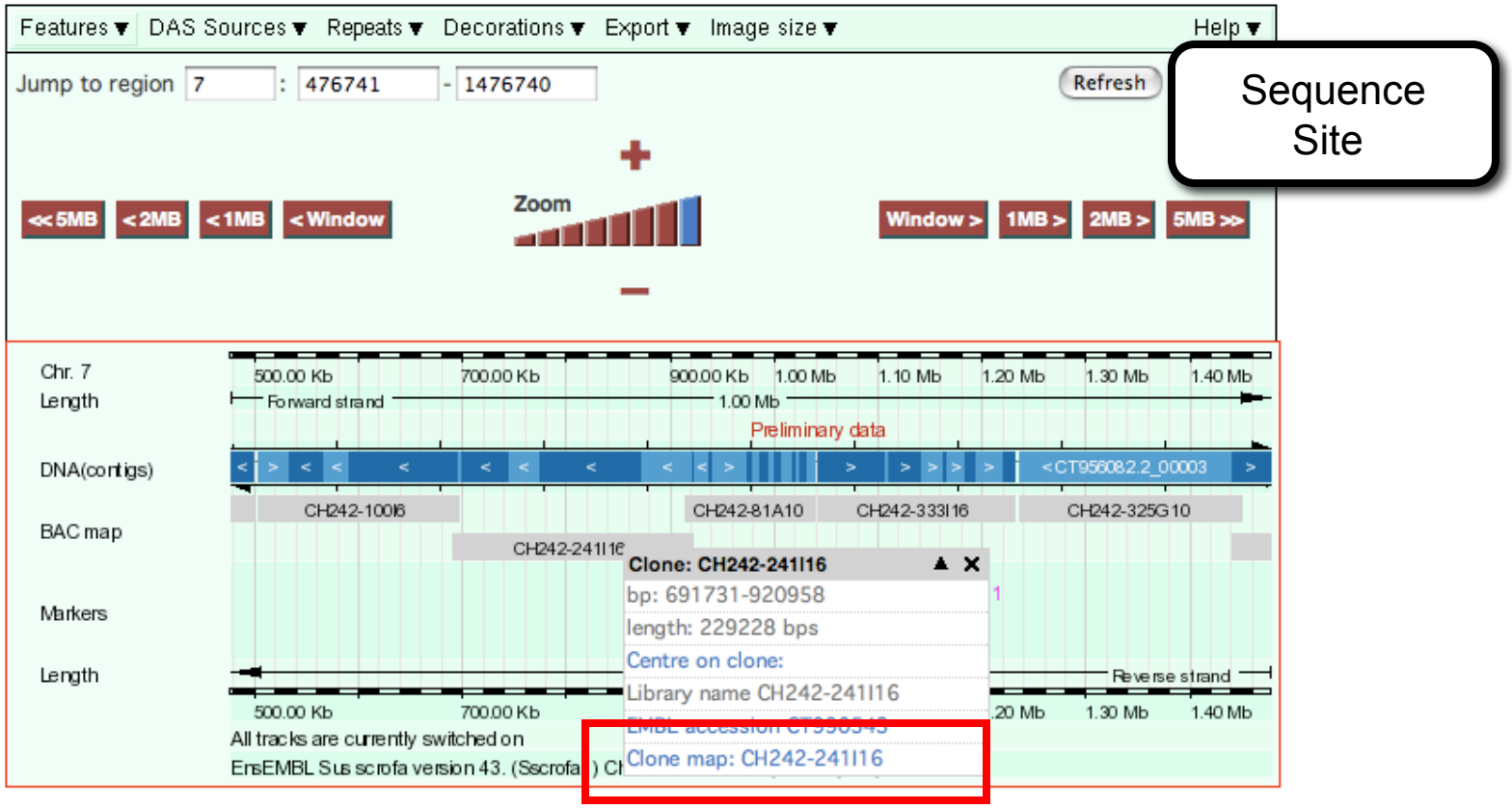
cDNAs in Regions not Covered by the MAP



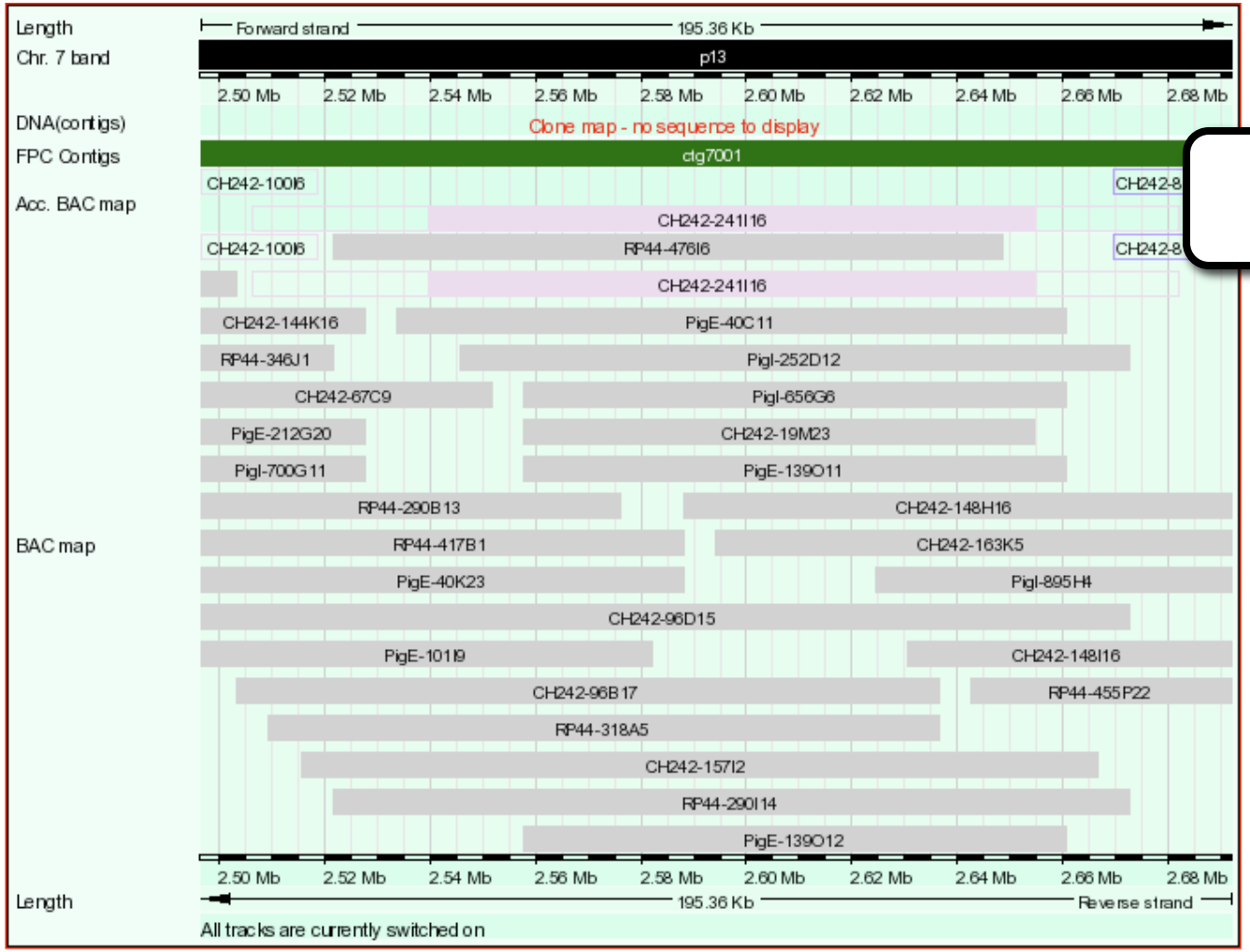
Tilepath


Features ▼
DAS Sources ▼
Repeats ▼
Decorations ▼
Export ▼
Image size ▼
Help ▼
Jump to region
7
476741
1476740
Refresh
Zoom
<< 5MB
< 2MB
< 1MB
< Window
Window >
1MB >
2MB >
5MB >>
Chr. 7
Length
500.00 Kb
700.00 Kb
900.00 Kb
1.00 Mb
1.10 Mb
1.20 Mb
1.30 Mb
1.40 Mb
Forward strand
1.00 Mb
Preliminary data
DNA(contigs)
CT956082.2_00003
BAC map
CH242-100I6
CH242-81A10
CH242-333I16
CH242-325G10
CH242-241I16
Markers
MARC_26445-26446:10306555521:1
PMC155124P3
311875963
Length
1.00 Mb
Reverse strand
All tracks are currently switched on
EnsEMBL Sus scrofa version 43. (Sscrofa1) Chromosome 7 476,741 - 1,476,740

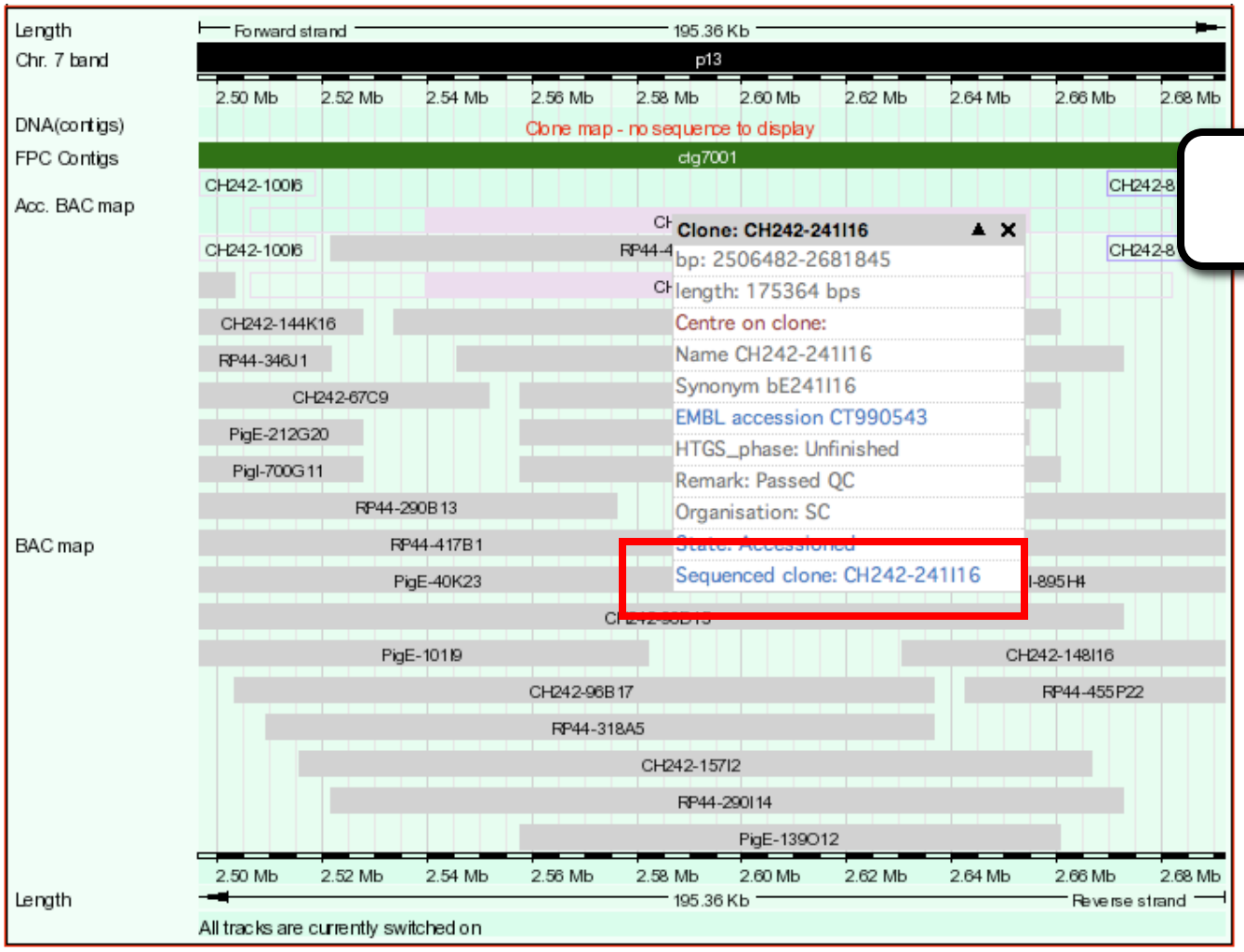
From the Sequence to the Map

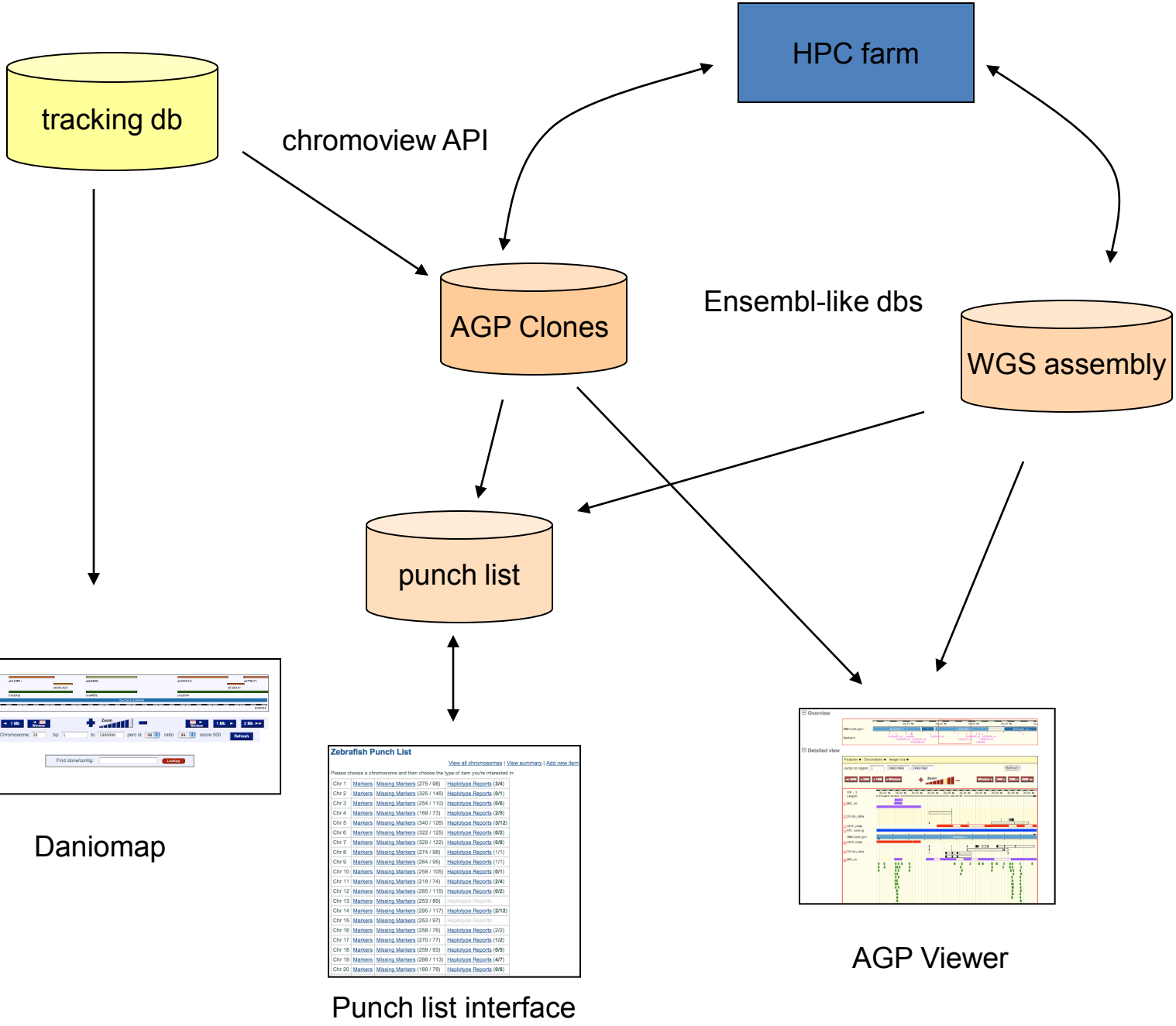


From the Sequence to the Map



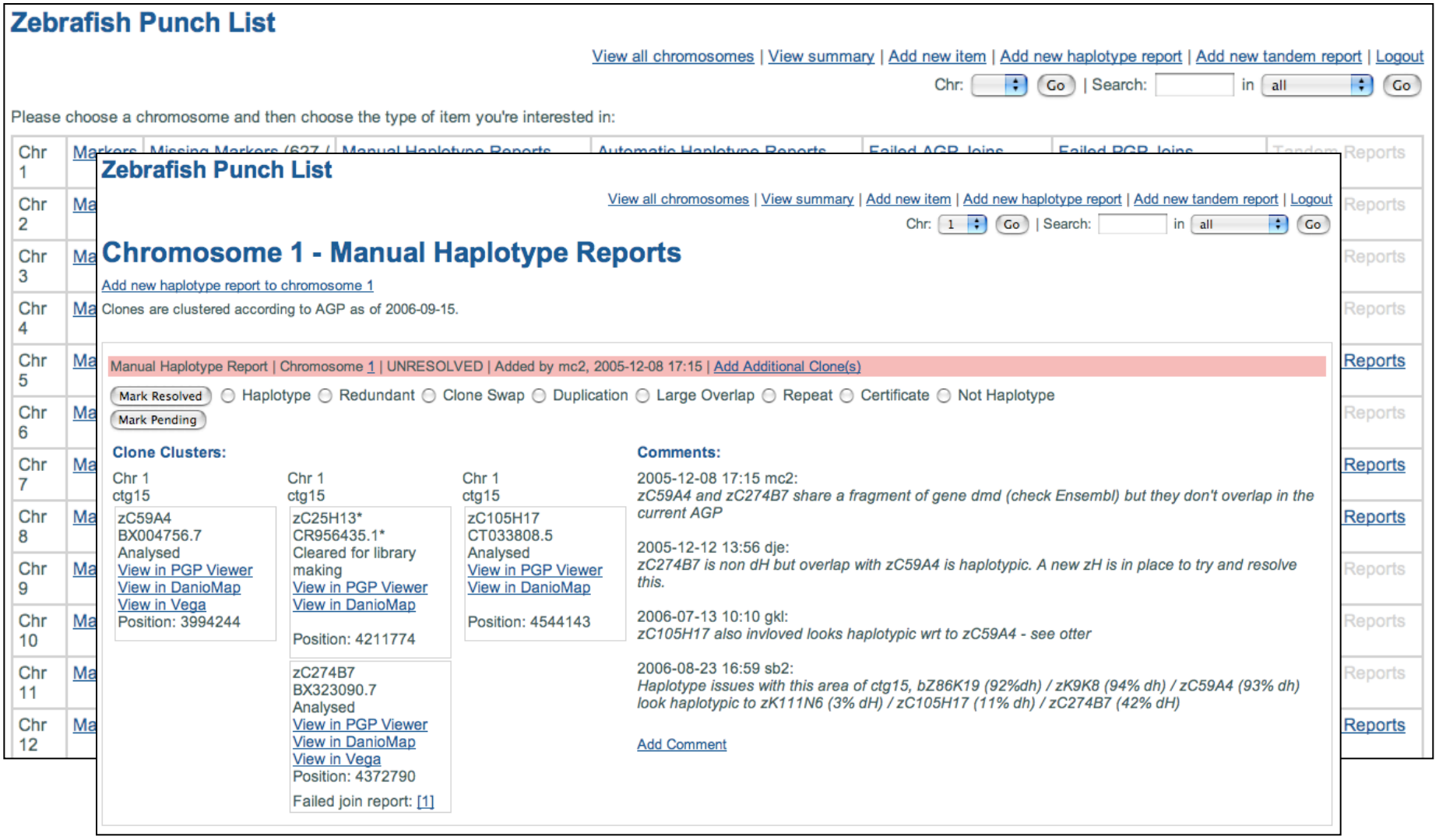
From the Map to the Sequence




HPC farm
tracking db
chromoview API
AGP Clones
Ensembl-like dbs
WGS assembly
punch list
DanioMap
Daniomap
Zebrafish Punch List
Punch list interface
AGP Viewer

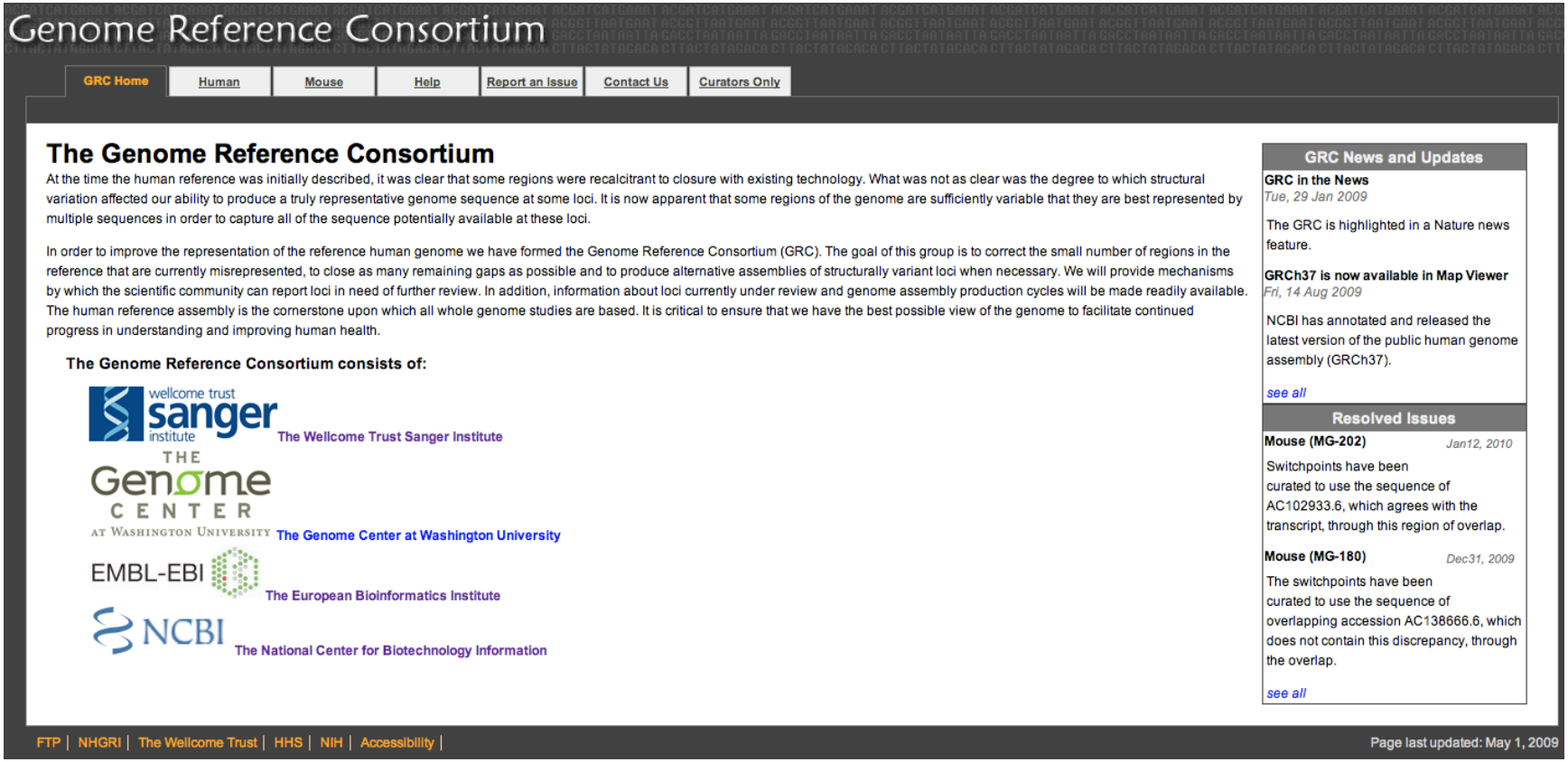
```
$slice=$slice_adaptor->fetch_by_region("chromosome",20);

my $clone_projs = $slice->project("clone");

foreach my $clone_proj (@$clone_projs){

    #get all bacends for $clone_proj


}
```

Punch Lists


Zebrafish Punch List
View all chromosomes | View summary | Add new item | Add new haplotype report | Add new tandem report | Logout
Chr: Go | Search: in all Go
Please choose a chromosome and then choose the type of item you're interested in:
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Reports
Zebrafish Punch List
View all chromosomes | View summary | Add new item | Add new haplotype report | Add new tandem report | Logout
Chr: 1 Go | Search: in all Go
Chromosome 1 - Manual Haplotype Reports
Add new haplotype report to chromosome 1
Clones are clustered according to AGP as of 2006-09-15.
Manual Haplotype Report | Chromosome 1 | UNRESOLVED | Added by mc2, 2005-12-08 17:15 | Add Additional Clone(s)
Mark Resolved
Haplotype
Redundant
Clone Swap
Duplication
Large Overlap
Repeat
Certificate
Not Haplotype
Mark Pending
Clone Clusters:
Chr 1
ctg15
zC59A4
BX004756.7
Analysed
View in PGP Viewer
View in DanioMap
View in Vega
Position: 3994244
Chr 1
ctg15
zC25H13*
CR956435.1*
Cleared for library making
View in PGP Viewer
View in DanioMap
Position: 4211774
zC274B7
BX323090.7
Analysed
View in PGP Viewer
View in DanioMap
View in Vega
Position: 4372790
Failed join report: [1]
Chr 1
ctg15
zC105H17
CT033808.5
Analysed
View in PGP Viewer
View in DanioMap
Position: 4544143
Comments:
2005-12-08 17:15 mc2:
zC59A4 and zC274B7 share a fragment of gene dmd (check Ensembl) but they don't overlap in the current AGP
2005-12-12 13:56 dje:
zC274B7 is non dH but overlap with zC59A4 is haplotypic. A new zH is in place to try and resolve this.
2006-07-13 10:10 gkl:
zC105H17 also invloved looks haplotypic wrt to zC59A4 - see otter
2006-08-23 16:59 sb2:
Haplotype issues with this area of ctg15, bZ86K19 (92%dh) / zK9K8 (94% dh) / zC59A4 (93% dh) look haplotypic to zK111N6 (3% dH) / zC105H17 (11% dh) / zC274B7 (42% dH)
Add Comment

Genome Reference Consortium

http://genomereference.org



Genome Reference Consortium

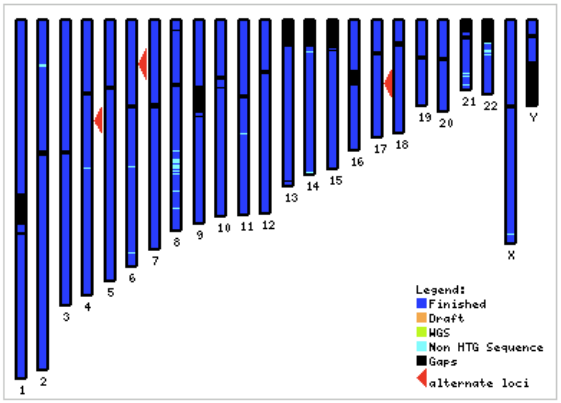
Genome Reference Consortium

GRC Home | Human | Mouse | Help | Report an Issue | Contact Us | Curators Only

Overview | Issues under Review | Assembly Data | Report a problem

Human Genome Overview

Information concerning continuing improvement of the human genome.



GRCh37: A graphical representation of the latest human assembly. The genome is colored with respect to the genomic component used to build the genome assembly at that location. The red triangles mark regions where alternate loci have been provided.

The most recent assembly for human is GRCh37. This is the first assembly produced by the GRC and is considered the next version of NCBI Build 36 (also known as hg18). Improvements in this assembly include:

- **Closure of 25 unspanned gaps found in Build 36**
- **Resolution of over 150 issues reported as problems in Build 36**
- **Addition of alternate loci for three complex regions, including the MHC region.**
- **Standardization of AGPs, including the addition of biological gap information.**

GRCh37 is a haploid assembly, constructed from multiple individuals and can be divided into a 'primary assembly' and a set of 'alternate loci'. The primary assembly represents the assembled chromosomes, plus any unlocalized or unplaced sequence that represent the non-redundant, haploid assembly. The alternate loci represent regions for which there is large scale variation and an alternate tilng path is available for this region. An example of such a region can be found at chromosome 17q21.31, often known as the MAPT locus. This region was described as carrying an inversion polymorphism (PMID: 15654335) and has been associated with various phenotypes (PMID: 16718704 ; PMID: 18628315). The version of this region in Build 36 was actually a mosaic of both haplotypes (as tracked in HG-77) and has been resolved in GRCh37 thanks to data described in Zody et al., 2008 (PMID: 19165922).

Information on alternate loci

Chromosome region with alternate loci	Length of region	Number of alternate contigs in region	View Region
UGT2B17 region (chr4:69,170,077-69,877,175)	707,099 bp	1 contig	view
MHC region (chr6: 28,477,797-33,448,354)	4,970,558 bp	7 contigs	view
MAPT region (chr17: 43,384,864-44,913,631)	1,528,768 bp	1 contig	view

GRC News and Updates

GRCh37 now available at UCSC
Fri, 8 May 2009

UCSC has released the latest version of the public human genome assembly.

GRCh37 now available in Pre! Ensembl
Thu, 9 Apr 2009

Ensembl has released the latest release of the public human genome assembly (GRCh37) on their Pre! site.

see all

Recently Resolved Human Issues

Human (HG-546) *May14, 2009*

The gap has been closed by adding CR394530.16 to the TPF

Human (HG-33) *May14, 2009*

CR812477 closes the gap

see all

References

Whole Genome Papers

The HGP Reference Assembly
The Venter Genome Assembly

Human Chromosome Papers

Chr1 Chr2 Chr3 Chr4 Chr5 Chr6
Chr7 Chr8 Chr9 Chr10 Chr11
Chr12 Chr13 Chr14 Chr15 Chr16
Chr17 Chr18 Chr19 Chr20 Chr21
Chr22 ChrX ChrY

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

Page last updated: April 22, 2009

Conclusions

- Integration of the data
  - speeds up project
  - validation
  - systematic identification of problematic areas

- Scientific Community
  - expertise in relevant areas of the genome
  - more and better eyes to look for issues
  - share ownership of the final product

- Quality
  - lasting resource
  - usability
  - architecture of the sequence