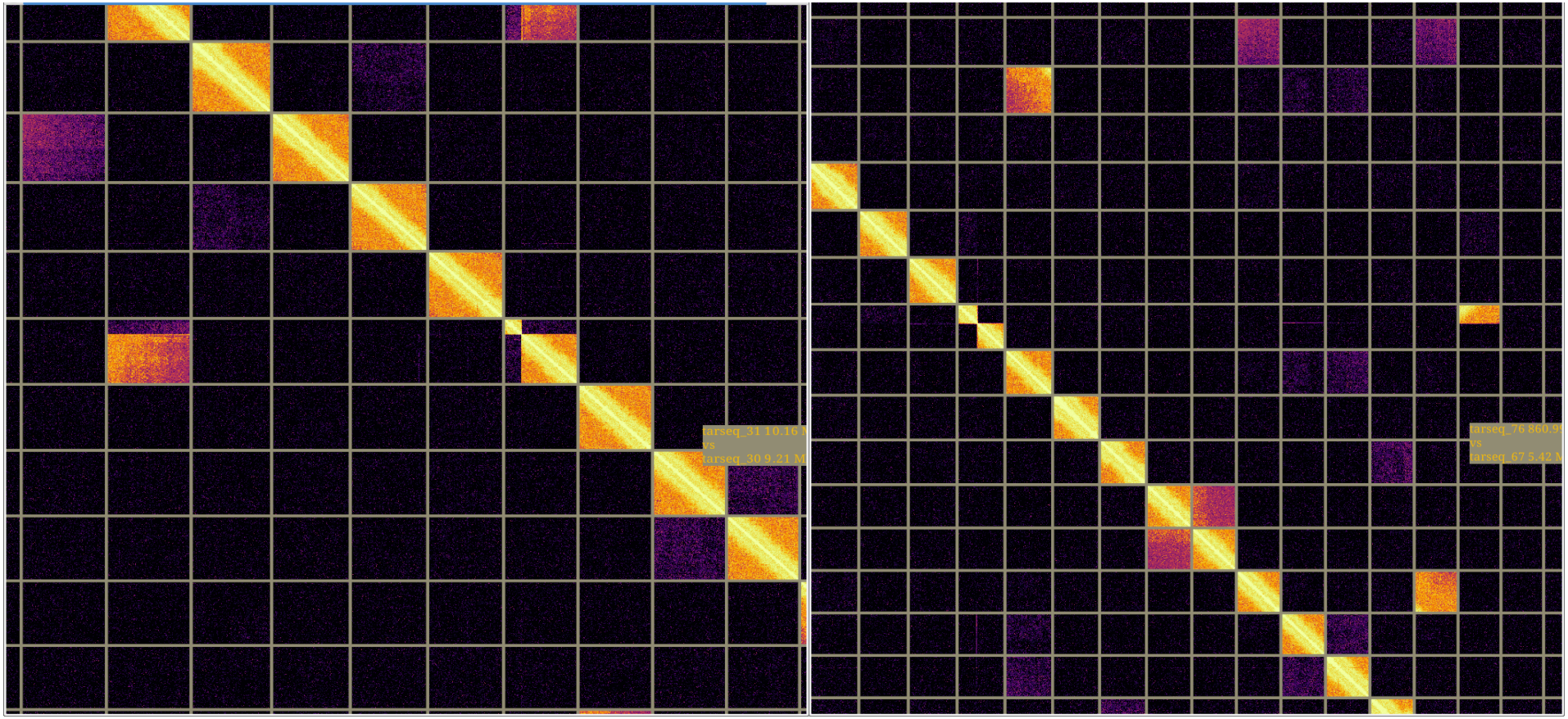
BreakHiC – Identification of Assembly Breakpoints by Screening Paired HiC Reads

Zemin Ning

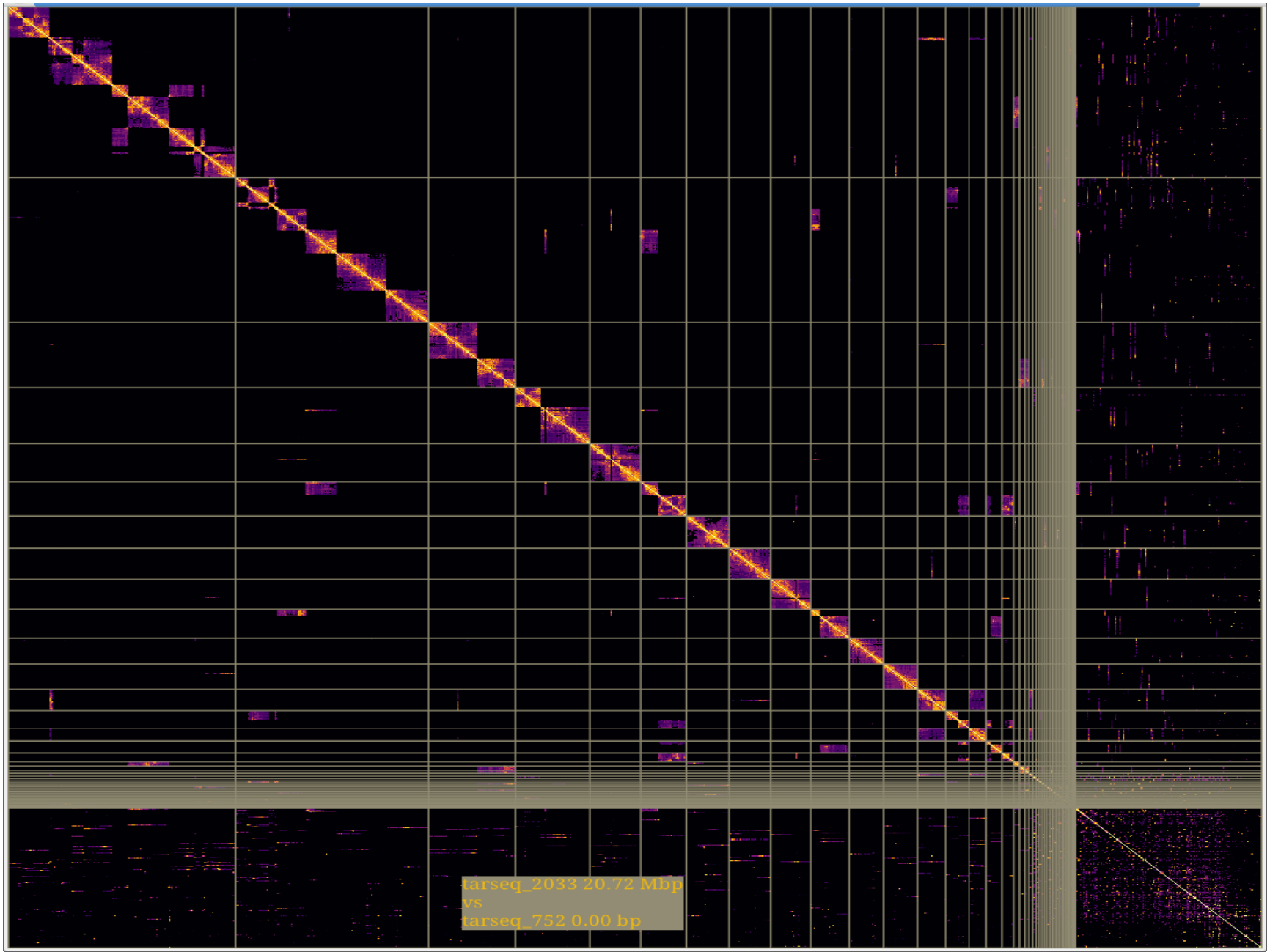
The Wellcome Sanger Institute

UK

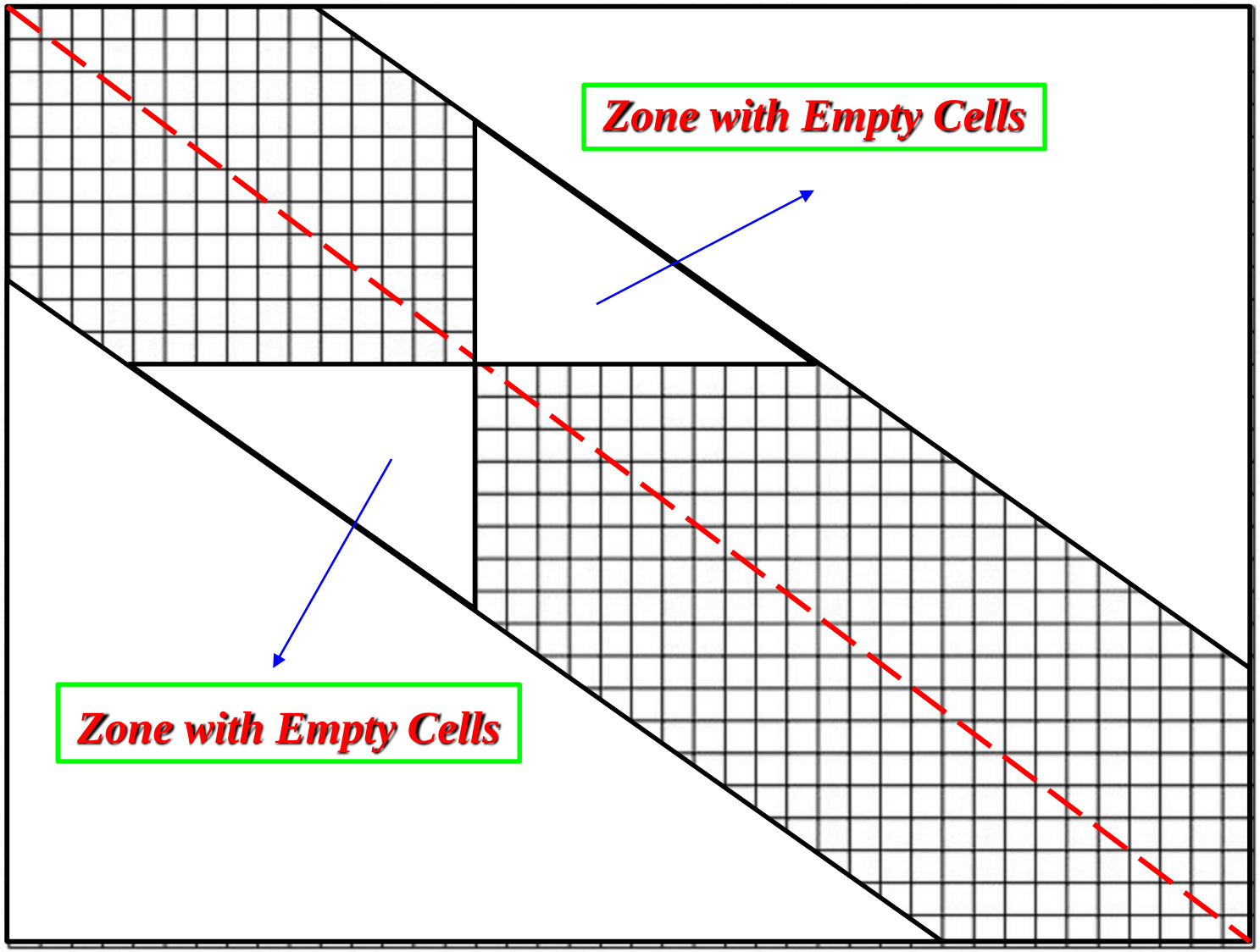
Mis-assembly Errors at Contigs



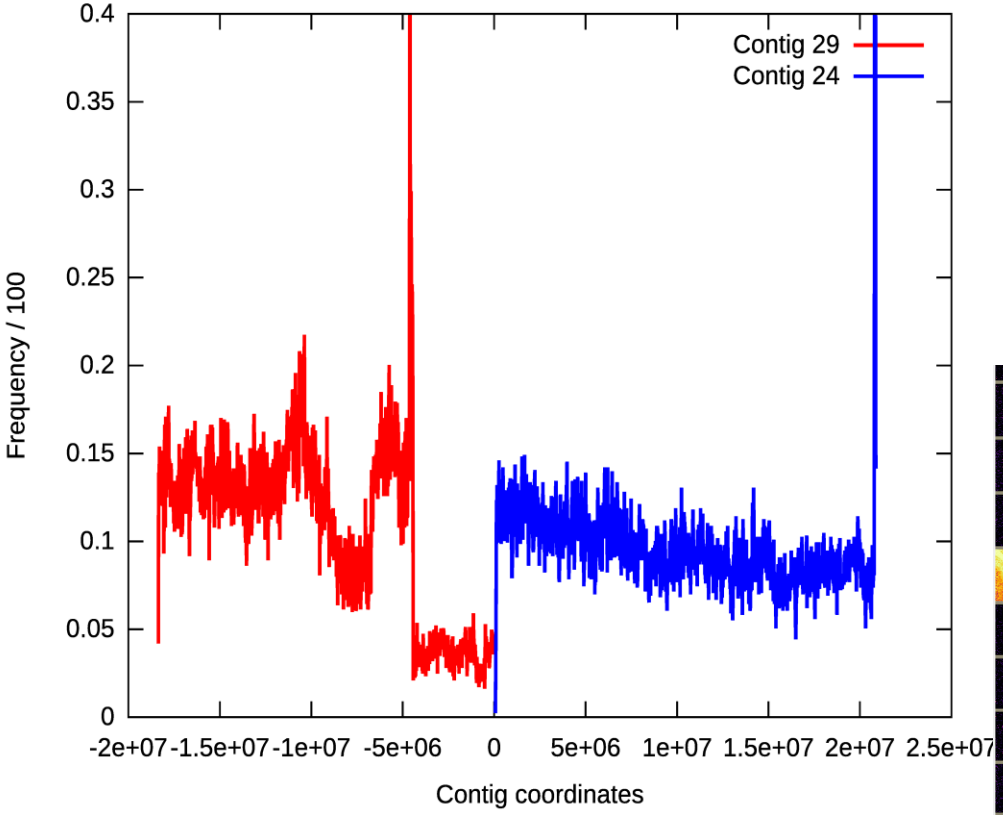
Mis-assembly Errors at Scaffolds

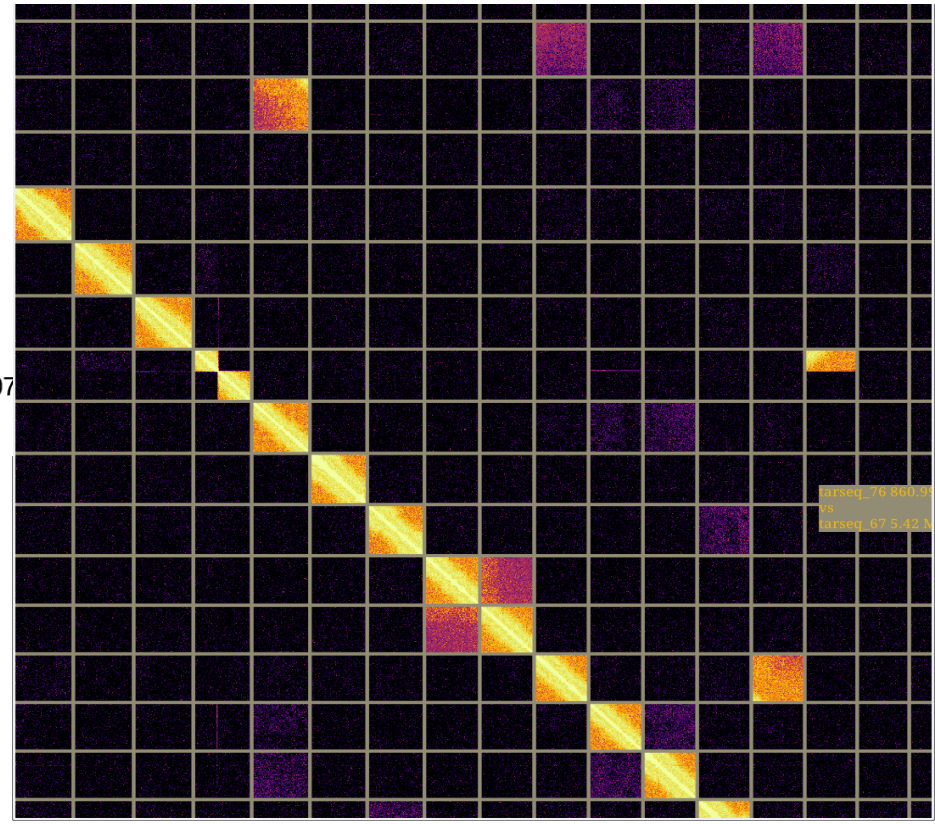


Grid Method to Check Empty Cells

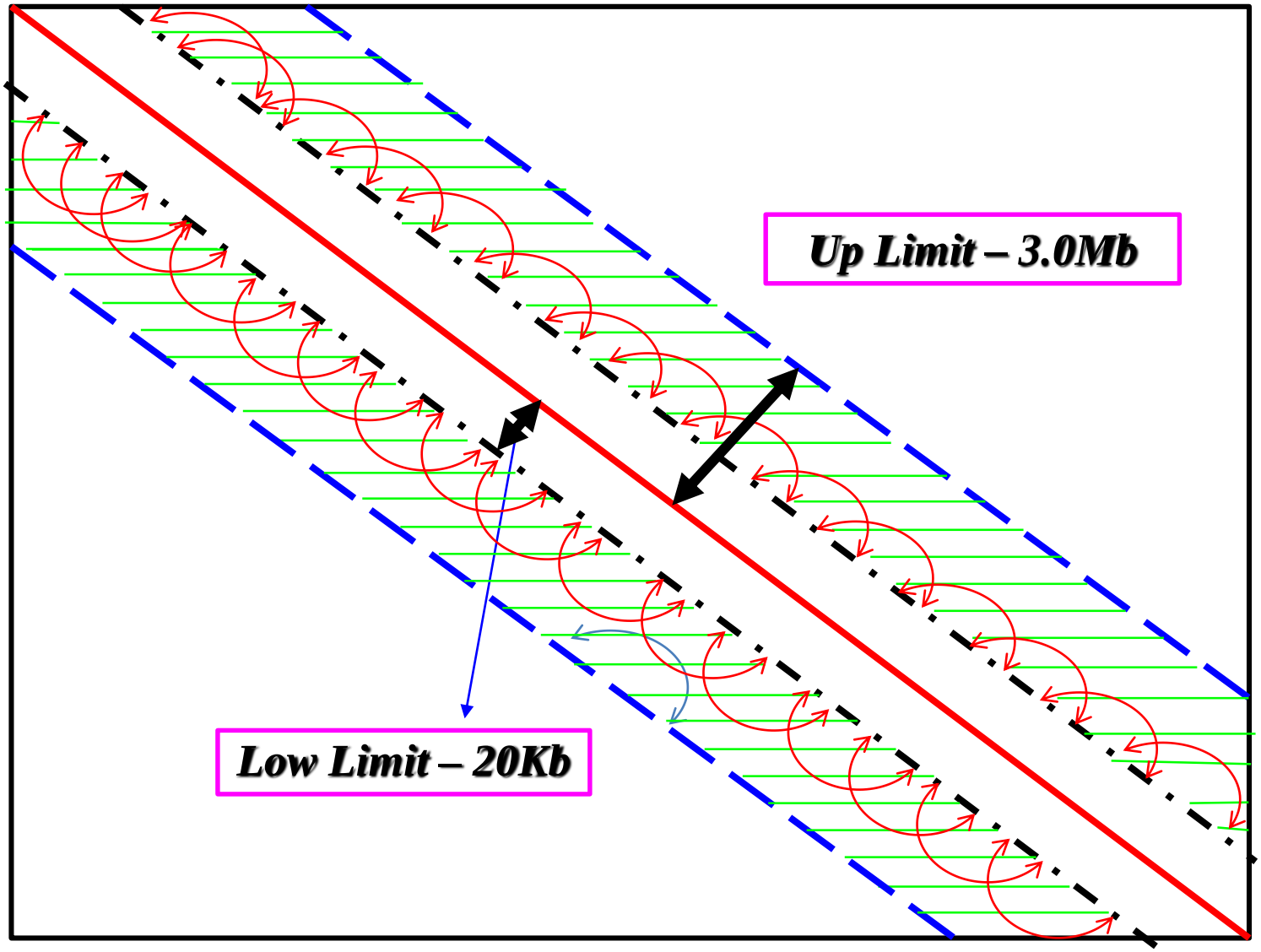


Number of Reads in Paired Contigs

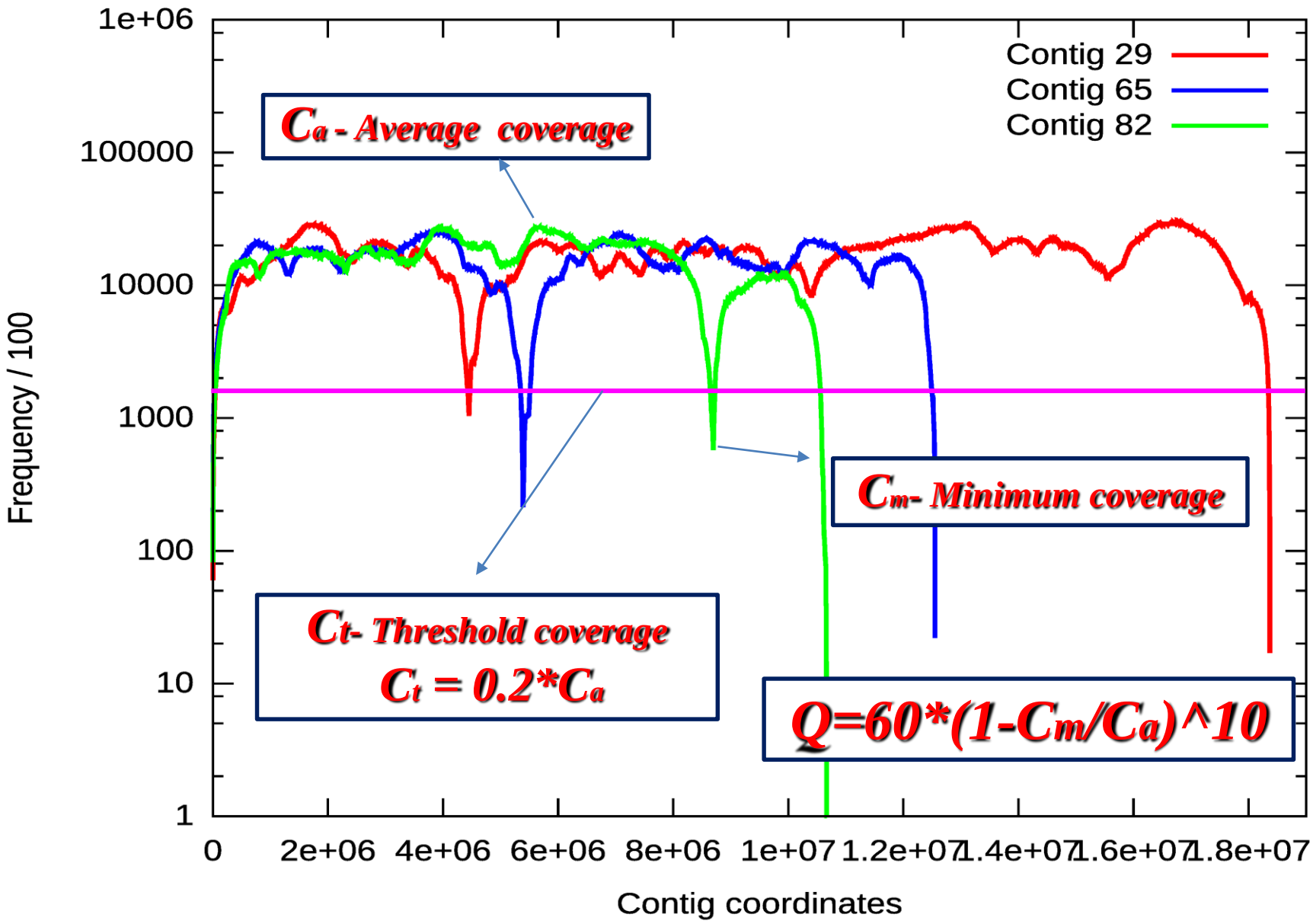




HiC Pair Link Coverage



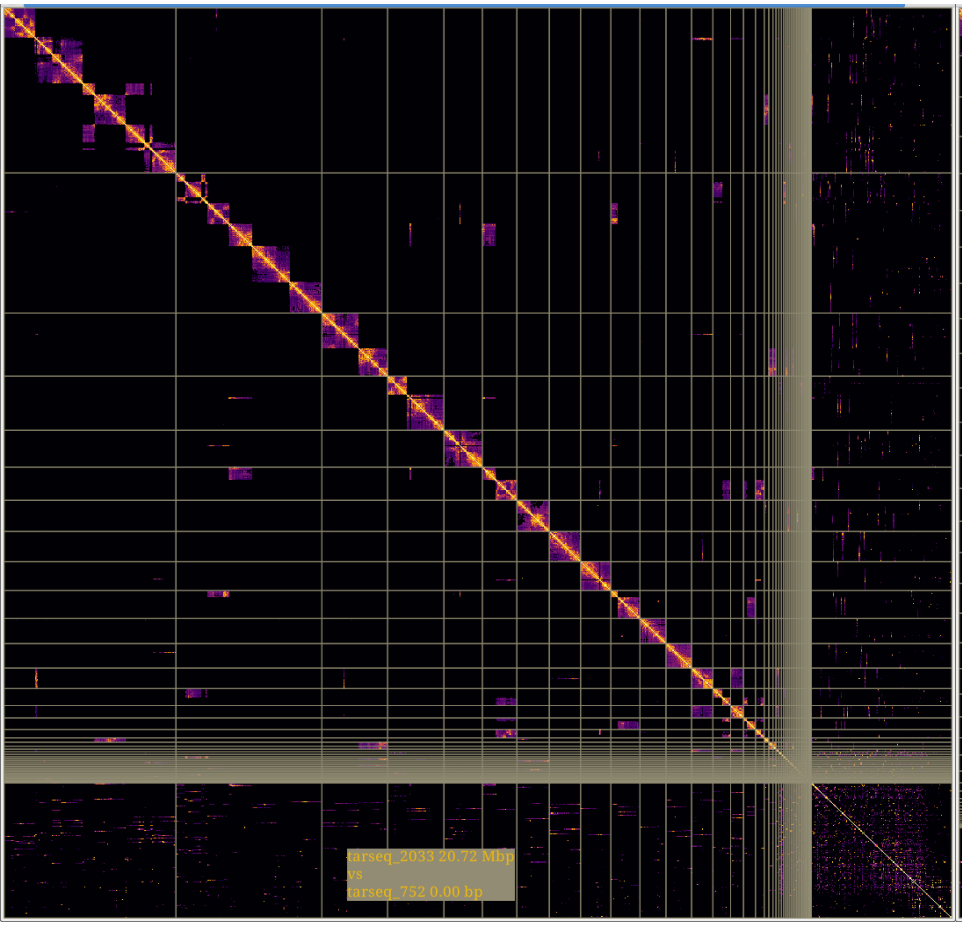
Coverage of HiC Pair Links within Contigs

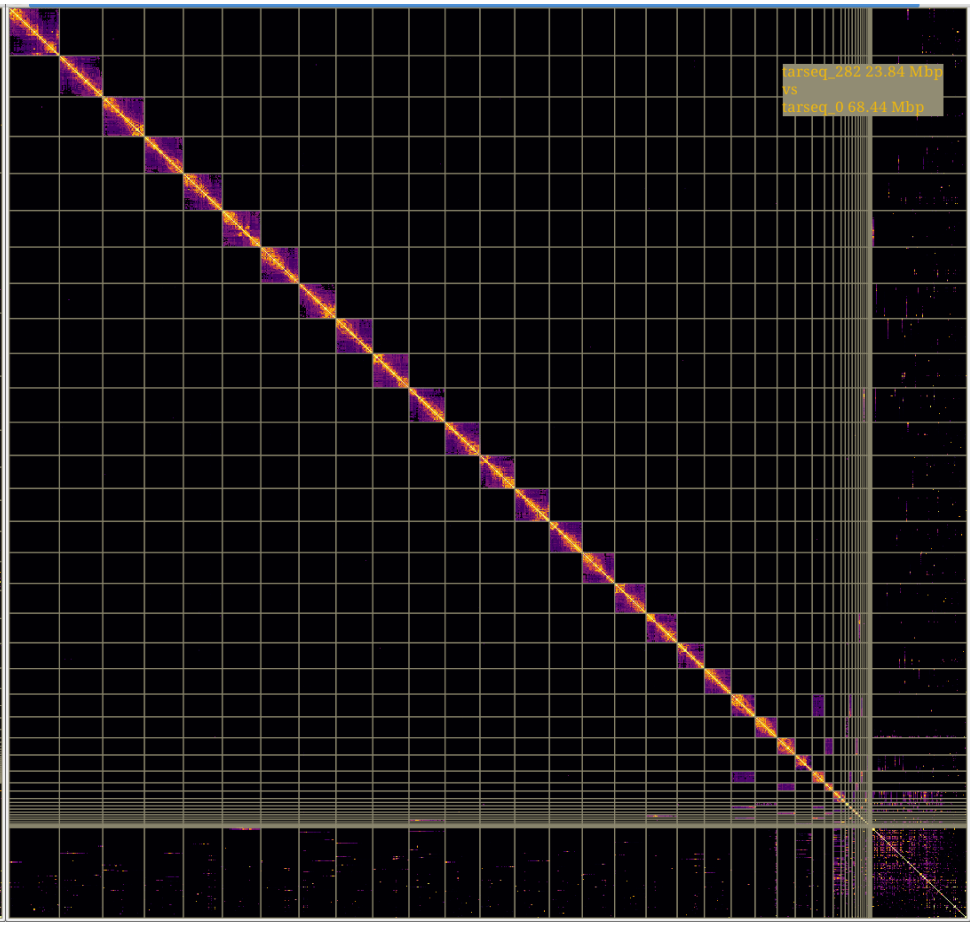


```
Break1: 70 547700 217464711 3156 5 0
Break1: 70 854600 217464711 3447 5 0
Break1: 70 1155600 217464711 2736 6 0
Break2: 70 119259859 217464711 3431 5 200
Break2: 117 1340007 72126970 389 38 200
Break2: 117 4228188 72126970 1157 12 200
Break2: 117 28301466 72126970 556 26 200
Break1: 118 46043000 76603601 1388 9 0
Break2: 118 60007606 76603601 179 60 200
Break2: 118 69709038 76603601 116 60 200

2 - Contig/scaffold index;
3 - Breakpoint offset;
4 - Contig/scaffold length;
5 - Average HiC coverage;
6 - Breakpoint likelihood value QV;
7 - Breakpoint nature: contig break (0); scaffold break or break at a gap (200)

====================================
https://github.com/wtsi-hpag/scaffHiC
====================================

$ git clone  https://github.com/wtsi-hpag/scaffhic.git
$ cd scaffhic
$ ./install.sh

/full/path/breakhic -nodes 30 -grid 100 -fq1 GM12878-HiC_1.fastq.gz -fq2 GM12878-HiC_2.fastq.gz
                    genome_assembly.fasta genome-break.fasta > try.out

/full/path/breakhic -nodes 2 -grid 100 -data /full/path/to/tmp_rununik_27152/align.dat
                    genome_assembly.fasta genome-break.fasta > try.out
```

***Output File* .brk**
and
Pipeline Download

Zfish – MAT-DHAB

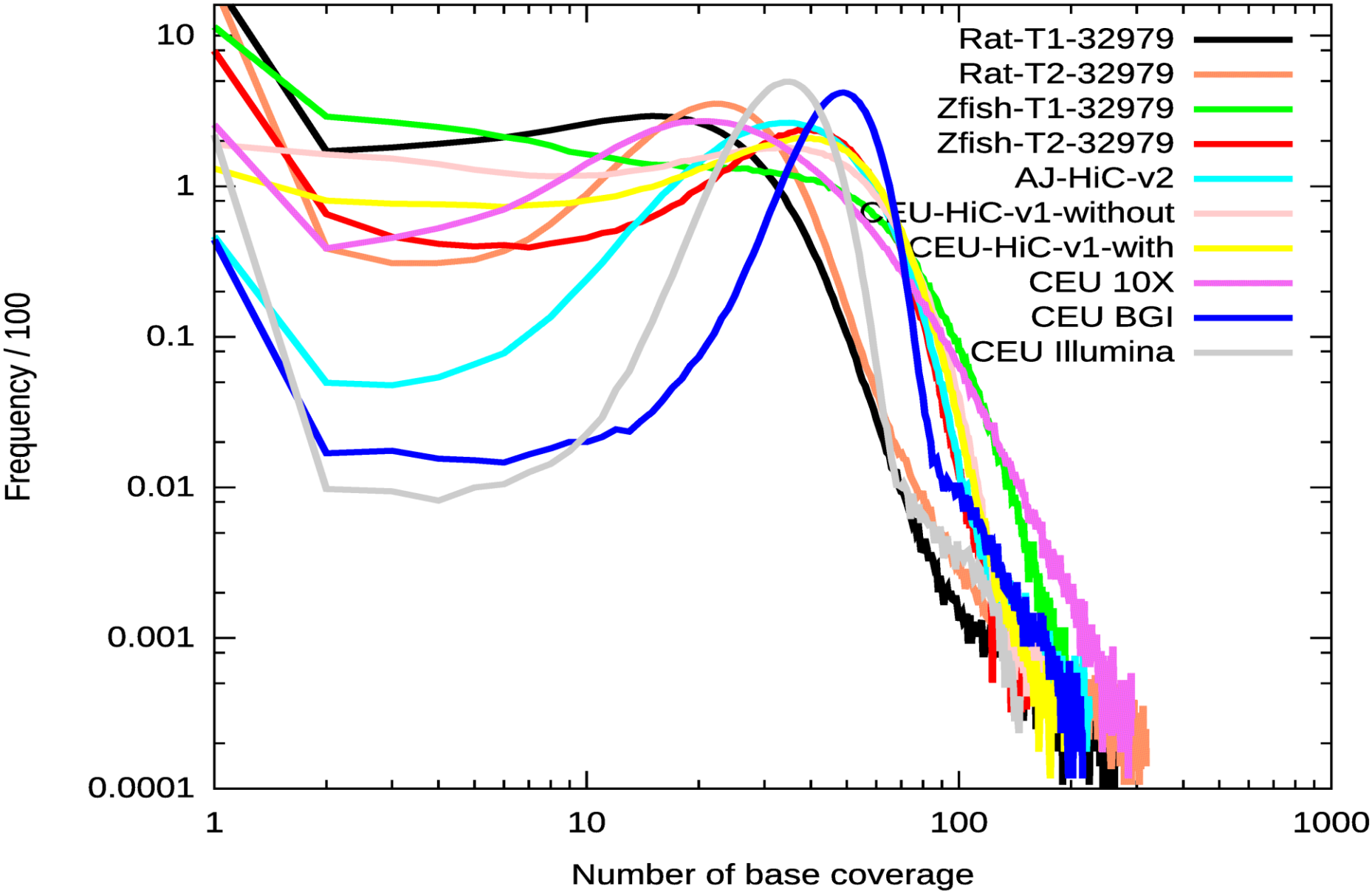




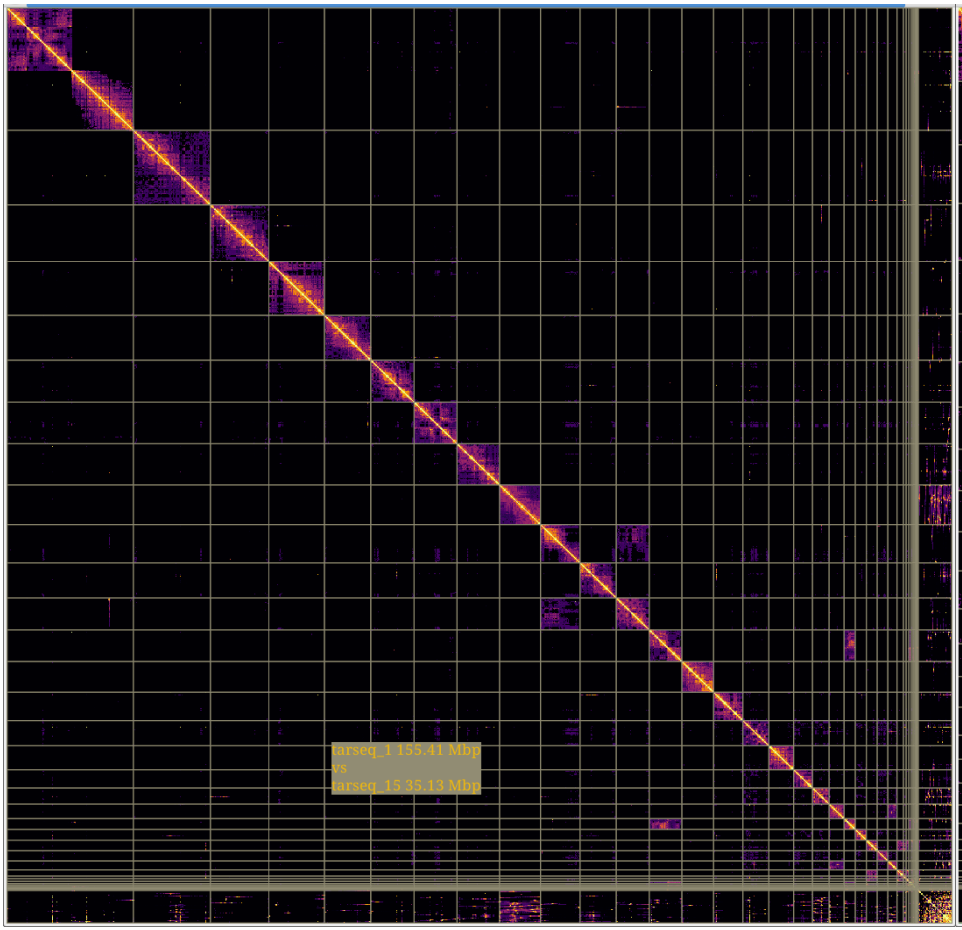
Without breakHiC

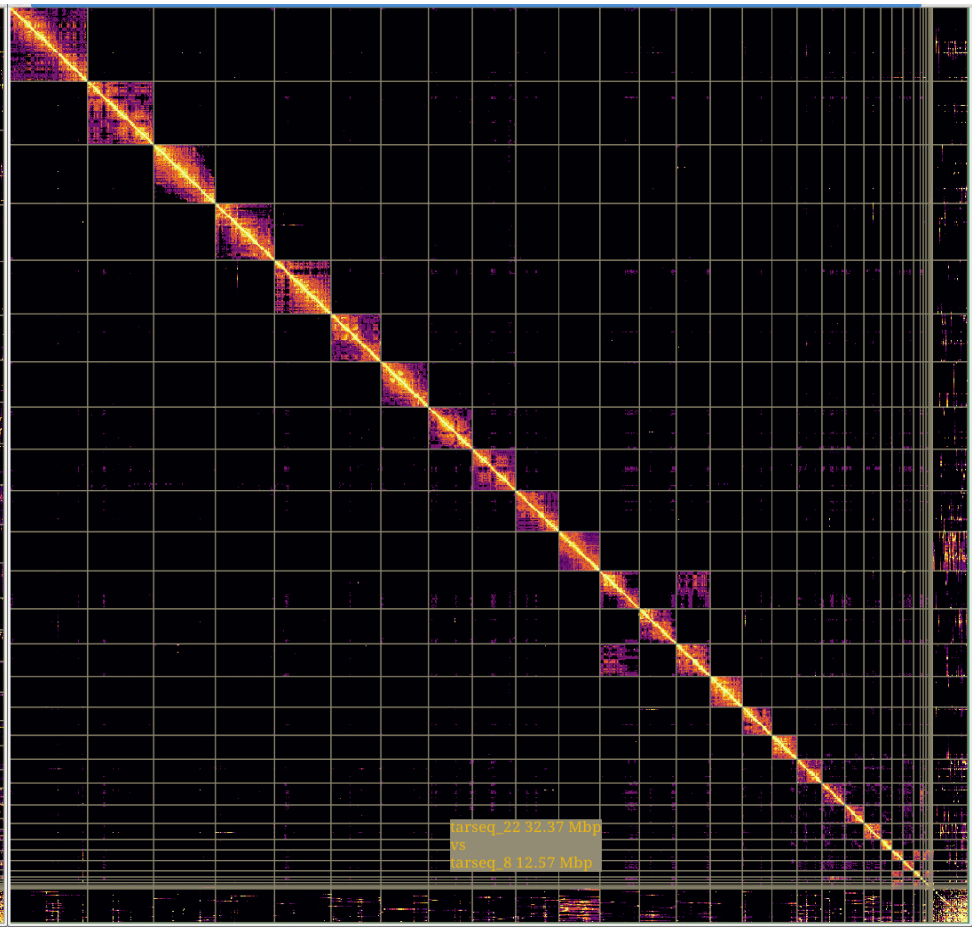
With breakHiC

Recently Sequenced Arima V2 on Rat and Zfish



Human Genome - Assembly

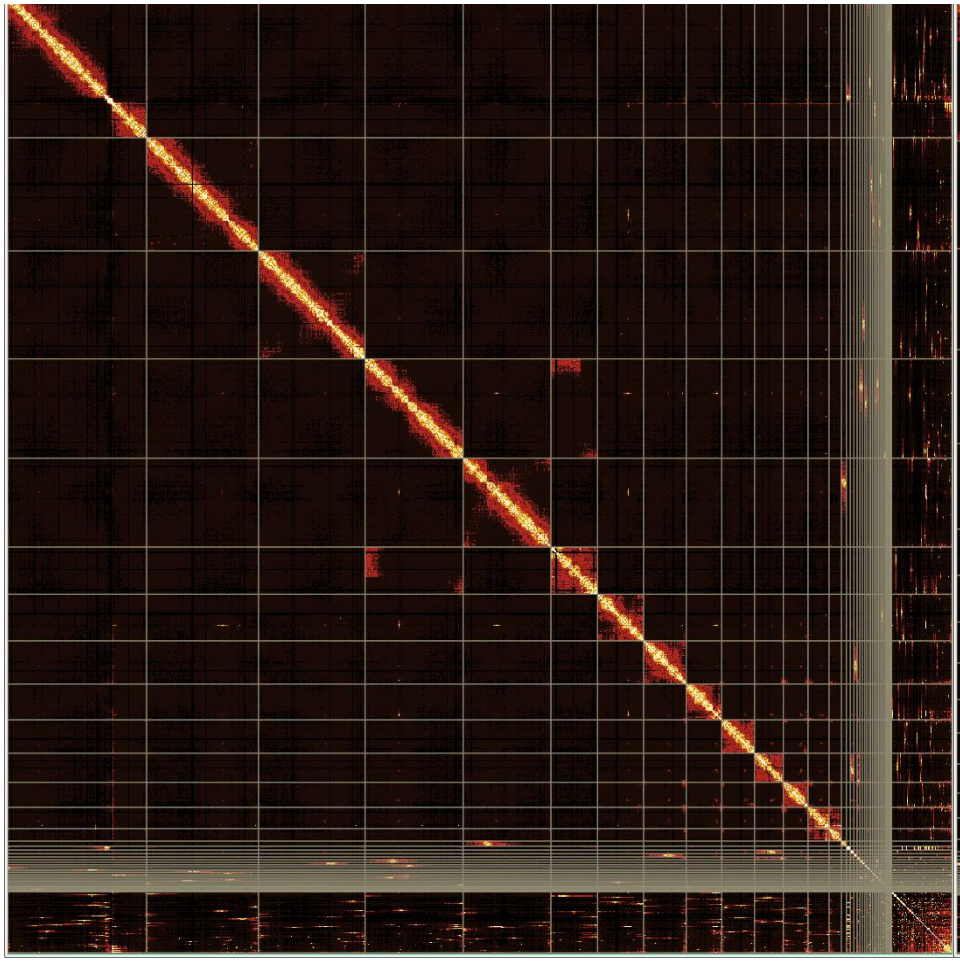




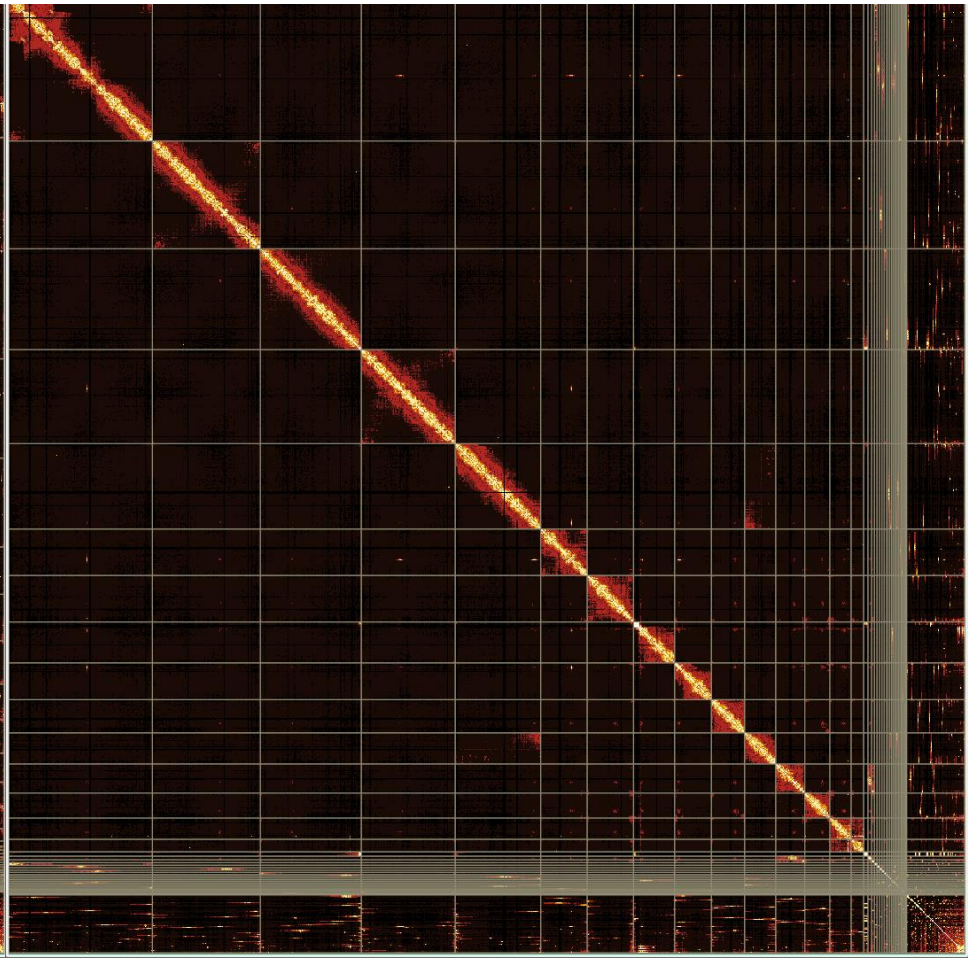
Without breakHiC

With breakHiC

Genome Assembly of Frog aRanTem1



Run 1



Run 2

Summary

- ***Reviewed three methods to identify assembly breakpoints***
- ***Grid method and distribution of reads in paired contigs have issues in finding exact breakpoints***
- ***HiC pair link coverage within contigs offers the best detection results***
- ***ScaffHiC with breakhic are working well for the assemblies with errors***
- ***More work is need to insert small contigs into established scaffolds to increase the ratio of chromosome assignment.***

Acknowledgements:

- *Mike Quail*
- *Dengfeng Guan*
- *Edward Harry*
- *Shane McCarthy*

Phase Genomics