Don't fear the REAPR: improved genome assembly with mate-pairs and long reads

ARTICLE

OPEN
doi:10.1038/nature13726

The genomic substrate for adaptive radiation in African cichlid fish

David Brawand[1,2]*, Catherine E. Wagner[3,4]*, Yang I. Li[2]*, Milan Malinsky[5,6], Irene Keller[4], Shaohua Fan[7], Oleg Simakov[7,8], Alvin Y. Ng[9], Zhi Wei Lim[9], Etienne Bezault[10], Jason Turner-Maier[1], Jeremy Johnson[1], Rosa Alcazar[11], Hyun Ji Noh[1], Pamela Russell[12], Bronwen Aken[6], Jessica Alföldi[1], Chris Amemiya[13], Naoual Azzouzi[14], Jean-François Baroiller[15], Frederique Barloy-Hubler[14], Aaron Berlin[1], Ryan Bloomquist[16], Karen L. Carleton[17], Matthew A. Conte[17], Helena D'Cotta[15], Orly Eshel[18], Leslie Gaffney[1], Francis Galibert[14], Hugo F. Gante[19], Sante Gnerre[1], Lucie Greuter[3,4], Richard Guyon[14], Natalie S. Haddad[16], Wilfried Haerty[2], Rayna M. Harris[20], Hans A. Hofmann[20], Thibaut Hourlier[6], Gideon Hulata[18], David B. Jaffe[1], Marcia Lara[1], Alison P. Lee[9], Iain MacCallum[1], Salome Mwaiko[3], Masato Nikaido[21], Hidenori Nishihara[21], Catherine Ozouf-Costaz[22], David J. Penman[23], Dariusz Przybylski[1], Michaelle Rakotomanga[14], Suzy C. P. Renn[10], Filipe J. Ribeiro[1], Micha Ron[18], Walter Salzburger[19], Luis Sanchez-Pulido[2], M. Emilia Santos[19], Steve Searle[6], Ted Sharpe[1], Ross Swofford[1], Frederick J. Tan[24], Louise Williams[1], Sarah Young[1], Shuangye Yin[1], Norihiro Okada[21,25], Thomas D. Kocher[17], Eric A. Miska[5], Eric S. Lander[1], Byrappa Venkatesh[9], Russell D. Fernald[11], Axel Meyer[7], Chris P. Ponting[2], J. Todd Streelman[16], Kerstin Lindblad-Toh[1,26], Ole Seehausen[3,4] & Federica Di Palma[1,27]

Cichlid fishes are famous for large, diverse and replicated adaptive radiations in the Great Lakes of East Africa. To understand the molecular mechanisms underlying cichlid phenotypic diversity, we sequenced the genomes and transcriptomes of five lineages of African cichlids: the Nile tilapia (*Oreochromis niloticus*), an ancestral lineage with low diversity; and four members of the East African lineage: *Neolamprologus brichardi/pulcher* (older radiation, Lake Tanganyika), *Metriaclima zebra* (recent radiation, Lake Malawi), *Pundamilia nyererei* (very recent radiation, Lake Victoria), and *Astatotilapia burtoni* (riverine species around Lake Tanganyika). We found an excess of gene duplications in the East African lineage compared to tilapia and other teleosts, an abundance of non-coding element divergence, accelerated coding sequence evolution, expression divergence associated with transposable element insertions, and regulation by novel microRNAs. In addition, we analysed sequence data from sixty individuals representing six closely related species from Lake Victoria, and show genome-wide diversifying selection on coding and regulatory variants, some of which were recruited from ancient polymorphisms. We conclude that a number of molecular mechanisms shaped East African cichlid genomes, and that amassing of standing variation during periods of relaxed purifying selection may have been important in facilitating subsequent evolutionary diversification.

Wide variation in the rates of diversification among lineages is a feature of evolution that has fascinated biologists since Darwin[1,2]. With approximately 2,000 known species, hundreds of which coexist in individual African lakes, cichlid fish are amongst the most striking examples of adaptive radiation, the phenomenon whereby a single lineage diversifies into many ecologically varied species in a short span of time[3] (Fig. 1). The largest radiations, which in Lakes Victoria, Malawi and Tanganyika, have generated between 250 (Tanganyika) and 500 (Malawi and Victoria) species per lake, took no more than 15,000 to 100,000 years for Victoria and less than 5 million years for Malawi[3–5], but 10–12 million years for Lake Tanganyika[6]. The radiations in Lake Victoria and Malawi thus display the highest sustained rates of speciation known to date in vertebrates[7]. The evolution of these lineages and their genomes has presumably been shaped by cycles of population expansion, fragmentation and contraction as lineages colonized lakes, diversified, collapsed when lakes dried up, and re-colonized lakes, and by episodic adaptation to a multitude of ecological niches coupled with strong sexual selection. Genetic diversity within lake radiations has been influenced by admixture following multiple colonization events and periodic infusions through hybridization[8,9].

Cichlid phenotypic diversity encompasses variation in behaviour, body shape, coloration and ecological specialization. The frequent occurrence of convergent evolution of similar ecotypes (Fig. 1) suggests a primary role of natural selection in shaping cichlid phenotypic diversity[10,11]. In addition, the importance of sexual selection is demonstrated by a profusion of exaggerated sexually dimorphic traits like male nuptial colour and elaborate bower building by males[3]. Ecological and sexual selection converge in the cichlid visual system, where trichromatic colour vision, eight different opsin genes and novel spherical lenses promote sensitivity in the highly dimensional visual world of clear-water lakes[12–14]. Rapidly evolving sex determination systems, often linked to male and female colour patterns, may also speed cichlid diversification[15,16]. Ecological, social and behavioural variation correlates with striking diversity in brain structures[17] that appears early in development[18].

Exceptional phenotypic variation, even among closely related species, makes cichlids different from most other fish groups, including those that share the same habitats with them but have not diversified as much, as well as those that have radiated into much smaller species flocks in northern temperate lakes[19]. However, how cichlids evolve in this exceptionally highly dimensional phenotype space remains unexplained.

We sequenced the genomes of five representative cichlid species from throughout the East African haplo-tilapiine lineage (Extended Data Fig 1a), which gave rise to all East African cichlid radiations. These five lineages diverged primarily through geographical isolation, and three of them subsequently underwent adaptive radiations in the three largest lakes of Africa (Fig. 1). Here we describe the comparative analyses of the five genomes coupled with an analysis of the genetic basis of species divergence in the Lake Victoria species flock to examine the genomic substrate for rapid evolutionary diversification.

Accelerated gene evolution

To assess whether accelerated sequence evolution was a general feature of East African cichlids, we annotated the genomes of all five cichlids (Extended Data Fig. 1a) and estimated the nonsynonymous/synonymous nucleotide substitution (d*N*/d*S*) ratio by sampling the concatenated alignments of all genes annotated with particular gene ontology (GO) terms. An elevated rate of nonsynonymous nucleotide substitutions can indicate accelerated evolution (either due to relaxed constraint or positive selection); this approach has been applied previously in the context of cichlid vision[13] and morphology[20,21]. We obtained significantly higher d*N*/d*S* ranks in *O. niloticus* (89 terms) compared to stickleback (11 terms), but considerably higher ranks still in the lineages of the East African radiation, haplochromines (299 terms) and *N. brichardi* (254 terms), (Extended Data Fig. 1b). In general, terms involved in morphological and developmental processes ranked significantly higher in haplochromines than in *O. niloticus* (P value = 0.036, Mann–Whitney U-test).

Amongst protein-coding genes with an increased number of nonsynonymous variants in haplochromines compared to *N. brichardi* and *O. niloticus*, two developmental genes, *nog2* and *bmpr1b*, emerged showing haplochromine-specific substitutions. This result is notable given that three genes, a ligand (*bmp4*)[21], a receptor (*bmpr1b*) and an antagonist (*nog2*) in the BMP pathway, all known to influence cichlid jaw morphology, show accelerated rates of protein evolution in haplochromine cichlids.

Of 22 candidate genes previously identified in teleost morphogenesis, vision and pigmentation, three are predicted to have undergone accelerated evolution in the common ancestors of the East African radiations suggesting a role in the diversification of cichlids: endothelin receptor type B1 (*ednrb1*) affects colour patterning[22] and perhaps pharyngeal jaw

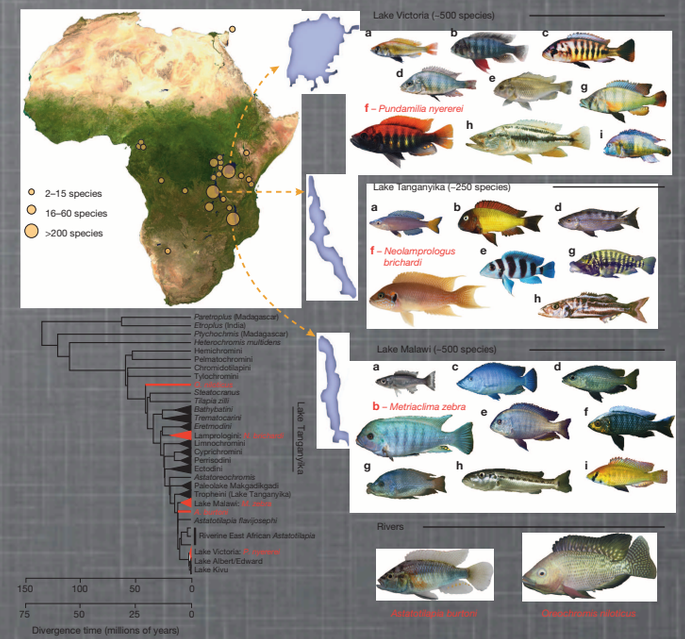
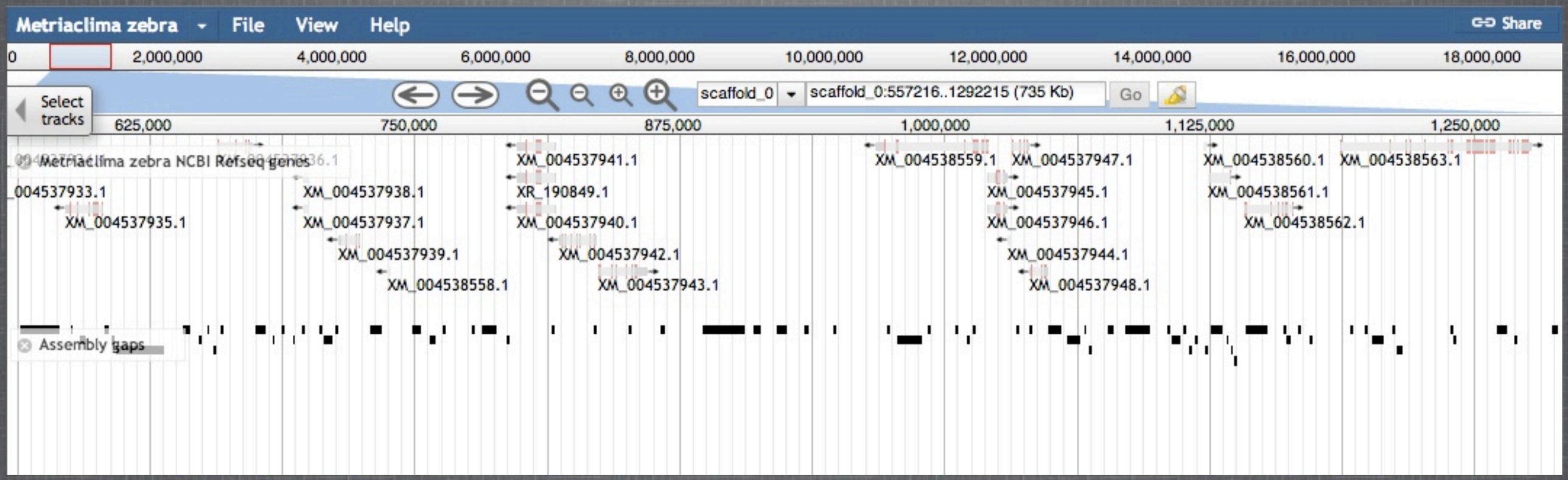


Figure 1 | The adaptive radiation of African cichlid fish. Top left, map of Africa showing lakes in which cichlid fish have radiated. Right, the five sequenced species: *Pundamilia nyererei* (endemic of Lake Victoria); *Neolamprologus brichardi* (endemic of Lake Tanganyika); *Metriaclima zebra* (endemic of Lake Malawi); *Oreochromis niloticus* (from rivers across northern Africa); *Astatotilapia burtoni* (from rivers connected to Lake Tanganyika). Major ecotypes are shown from each lake: **a**, pelagic zooplanktivore; **b**, rock-dwelling algae scraper; **c**, paedophage (absent from Lake Tanganyika); **d**, scale eater; **e**, snail crusher; **f**, reef-dwelling planktivore; **g**, lobe-lipped insect eater; **h**, pelagic piscivore; **i**, ancestral river-dweller also found in lakes (absent from Lake Tanganyika). Bottom left, phylogenetic tree illustrating relationships between the five sequenced species (red), major adaptive radiations and major river lineages. The tree is from ref. 4, pruned to the major lineages. Upper timescale (4), lower timescale (32). Photos by Ad Konings (Tanganyika **a**, **b**, **d**, **e**, **g**, **h**; Malawi **a**, **c**, **d**, **e**, **f**, **g**, **h**, **i**), O.S. (Victoria **a**–**g**, **i**; Malawi **b**), Frans Witte (Victoria **h**), W.S. (Tanganyika **f**), Oliver Selz (Victoria **f**, *A. burtoni*), Marcel Haesler (*O. niloticus*).

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [2]MRC Functional Genomics Unit, University of Oxford, Oxford OX1 3QX, UK. [3]Department of Fish Ecology and Evolution, Eawag Swiss Federal Institute of Aquatic Science and Technology, Center for Ecology, Evolution & Biogeochemistry, CH-6047 Kastanienbaum, Switzerland. [4]Division of Aquatic Ecology, Institute of Ecology & Evolution, University of Bern, CH-3012 Bern, Switzerland. [5]Gurdon Institute, Cambridge CB2 1QN, UK. [6]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [7]Department of Biology, University of Konstanz, D-78457 Konstanz, Germany. [8]European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [9]Institute of Molecular and Cell Biology, A*STAR, 138673 Singapore. [10]Department of Biology, Reed College, Portland, Oregon 97202, USA. [11]Biology Department, Stanford University, Stanford, California 94305-5020, USA. [12]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA. [13]Benaroya Research Institute at Virginia Mason, Seattle, Washington 98101, USA. [14]Institut Génétique et Développement, CNRS/University of Rennes, 35043 Rennes, France. [15]CIRAD, Campus International de Baillarguet, TA B-110/A, 34398 Montpellier cedex 5, France. [16]School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332-0230, USA. [17]Department of Biology, University of Maryland, College Park, Maryland 20742, USA. [18]Animal Genetics, Institute of Animal Science, ARO, The Volcani Center, Bet-Dagan, 50250 Israel. [19]Zoological Institute, University of Basel, CH-4051 Basel, Switzerland. [20]Department of Integrative Biology, Center for Computational Biology and Bioinformatics; The University of Texas at Austin, Austin, Texas 78712, USA. [21]Department of Biological Sciences, Tokyo Institute of Technology, Tokyo, 226-8501 Yokohama, Japan. [22]Systématique, Adaptation, Evolution, National Museum of Natural History, 75005 Paris, France. [23]Institute of Aquaculture, University of Stirling, Stirling FK9 4LA, UK. [24]Carnegie Institution of Washington, Department of Embryology, 3520 San Martin Drive Baltimore, Maryland 21218, USA. [25]National Cheng Kung University, Tainan City, 704 Taiwan. [26]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, 751 23 Uppsala, Sweden. [27]Vertebrate and Health Genomics, The Genome Analysis Centre, Norwich NR18 7UH, UK.
*These authors contributed equally to this work.

©2014 Macmillan Publishers Limited. All rights reserved

Rationale



- gaps
- missing gene duplicates
- mis/over-assemblies

Hunt *et al. Genome Biology* 2013, **14**:R47
http://genomebiology.com/2013/14/5/R47

Genome Biology

SOFTWARE Open Access

REAPR: a universal tool for genome assembly evaluation

Martin Hunt[1], Taisei Kikuchi[1,2], Mandy Sanders[1], Chris Newbold[1,3], Matthew Berriman[1] and Thomas D Otto[1*]

Abstract

Methods to reliably assess the accuracy of genome sequence data are lacking. Currently completeness is only described qualitatively and mis-assemblies are overlooked. Here we present REAPR, a tool that precisely identifies errors in genome assemblies without the need for a reference sequence. We have validated REAPR on complete genomes or *de novo* assemblies from bacteria, malaria and *Caenorhabditis elegans*, and demonstrate that 86% and 82% of the human and mouse reference genomes are error-free, respectively. When applied to an ongoing genome project, REAPR provides corrected assembly statistics allowing the quantitative comparison of multiple assemblies. REAPR is available at http://www.sanger.ac.uk/resources/software/reapr/.

Keywords: Genome assembly, validation, evaluation

Hunt *et al. Genome Biology* 2013, **14**:R47
http://genomebiology.com/2013/14/5/R47

Genome Biology

SOFTWARE Open Access

REAPR: a universal tool for genome assembly evaluation

Martin Hunt[1], Taisei Kikuchi[1,2], Mandy Sanders[1], Chris Newbold[1,3], Matthew Berriman[1] and Thomas D Otto[1*]

Abstract

Methods to reliably asse
described qualitatively a
errors in genome assem
genomes or *de novo* ass
82% of the human and
genome project, REAPR
assemblies. REAPR is ava

mpleteness is only
at precisely identifies
REAPR on complete
onstrate that 86% and
to an ongoing
parison of multiple

Keywords: Genome ass

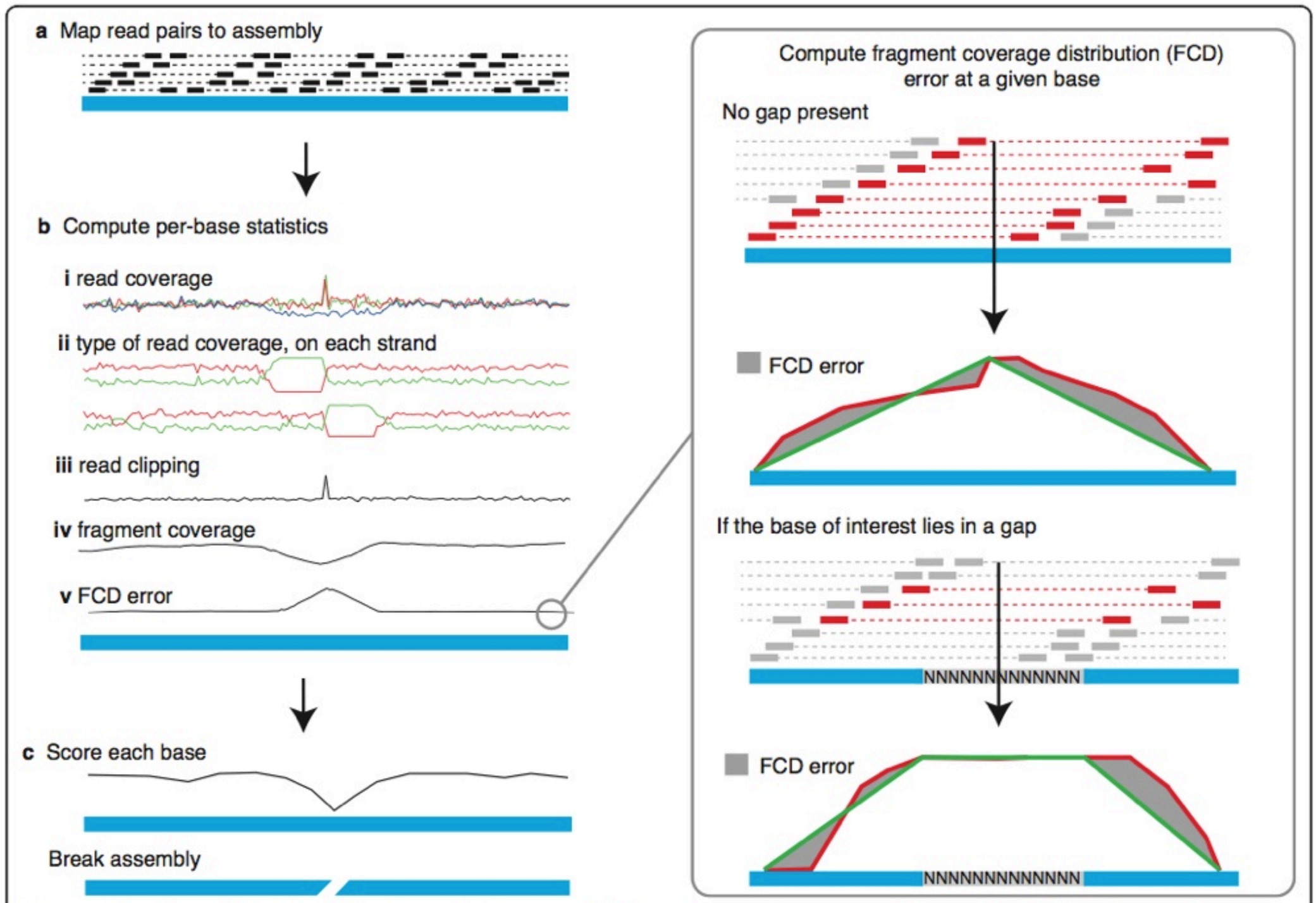


Figure 1 Overview of the REAPR pipeline. (**a**) The input is a BAM file of read pairs mapped to the assembly. (**b**) Statistics are calculated at each base of the genome: (**i**) Read coverage per strand, and any perfect and uniquely mapped read coverage is incorporated; (**ii**) The type of read coverage on the forward (upper plots) and reverse (lower plots) strand: proportion of reads that are properly paired (red), orphaned (green), and in the wrong orientation or exceed the fragment size range (not shown); (**iii**) The number of reads soft-clipped at each base; (**iv**) The fragment coverage, determined by the properly paired reads; (**v**) FCD error, taking into account the presence of a gap. Boxed are: FCD calculation at a given base. The fragments covering that base, shown in red, are used to construct a fragment depth plot (red). The FCD error is the area (grey) between the observed (red) plot and ideal plot (green). Since no read can map to a gap in the assembly, the calculation is corrected when a gap is present. (**c**) The statistics at each base are used independently to assign a score to each base of the assembly and also to break the assembly at scaffolding errors.

Table 1 A summary of REAPR results on a range of genome sequences.

Genome assembly	Total length (Mb)	Gaps (*n*)	Total gap length (bp)	Original N50 (Mb)	Corrected N50[b] (Mb)	Scaffold errors[a]			
						Called by REAPR	False +ve	False -ve	Error-free bases (%)
S. aureus TW20 k71	3.0	31	249	0.2	0.2	18	2	0	98.2
S. aureus, GAGE Velvet	2.9	128	17,688	0.8	0.2	24	0	1	89.5
P. falciparum de novo k55	23.8	11,636	2,638,349	0.4	0.3	56	1	8	81.2
P. falciparum v2.1.4	23.3	160	947	1.7	1.7	4	1	0	94.5
P. falciparum v3	23.3	0	0	1.7	1.7	NA	NA	NA	94.9
C. elegans WS228	100.3	0	0	17.5	17.5	NA	NA	NA	90.3
M. musculus GRCm38	2725.5	522	77,999,939	130.7	100.2	41	ND	ND	80.1
H. sapiens GRCh37	3095.7	360				234,350,278	155.3	146.4	6
ND	ND	79.1							

[a]Scaffold errors are not applicable (NA) when the assembly contains no gaps. Where a second genome sequence was unavailable for comparison, false-positives and false-negatives were not determined (ND).

[b]Corrected N50 refers to the N50 of the assembly after breaking the original assembly at breakpoints called by REAPR.

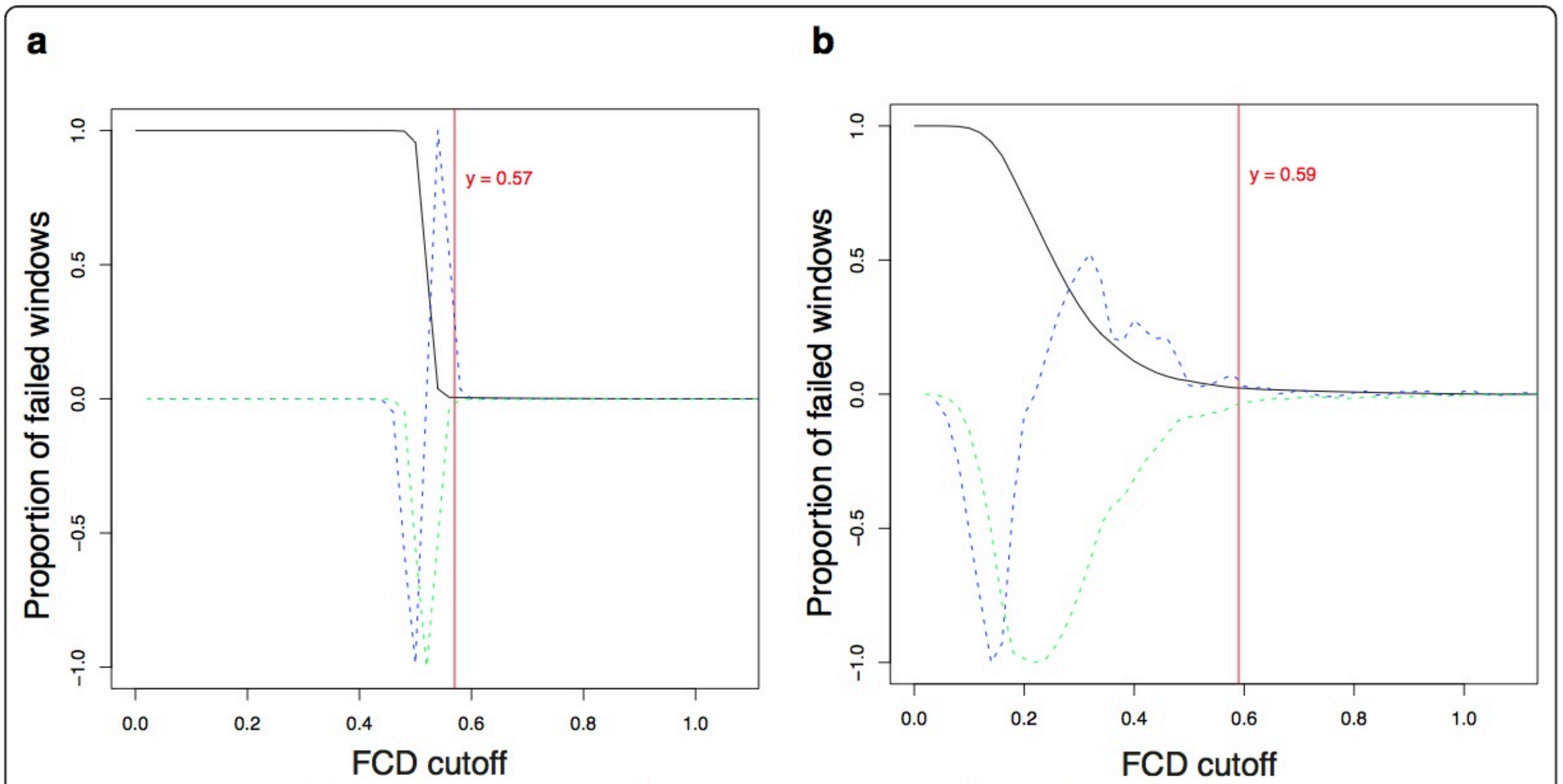
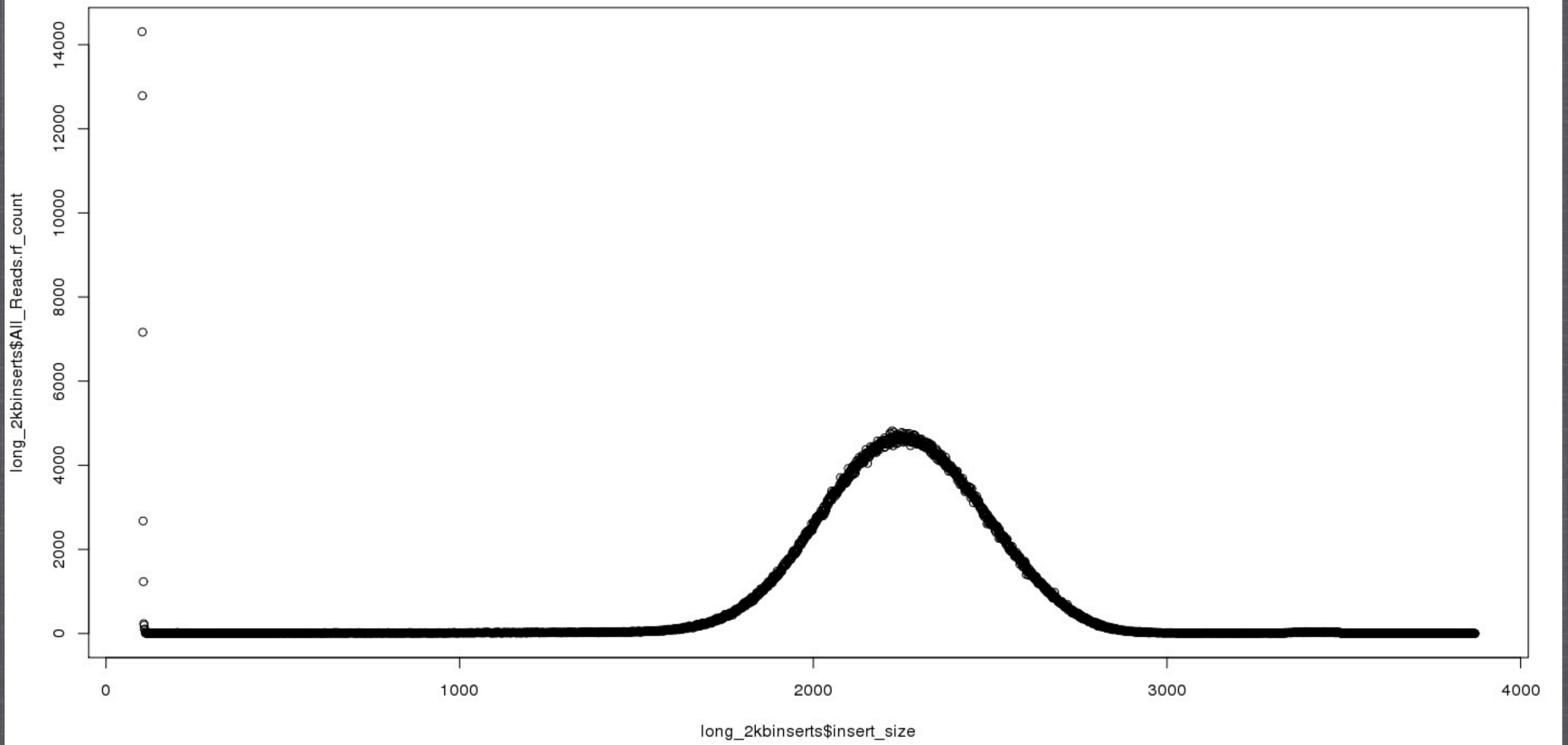


Figure 2 Automatic calculation of the FCD error cutoff. (**a**) *S. aureus de novo* assembly (*k*-mer of 71). (**b**) *P. falciparum de novo* assembly (*k*-mer of 55). In each plot, the black line shows the proportion of windows that would be called as an error for a range of cutoff values. The green and blue lines are the first and second derivatives of the black line, normalised to lie between -1 and 1. The vertical red line marks the FCD error cutoff, automatically determined by REAPR as the first FCD score corresponding to first and second derivatives ≥0.05.

Insert size distribution of Broad libraries (Assemblathon2 instructions)

Type	Name	Reads	Bases	Coverage
fragment	Solexa-38739	597610332	60358643532	60
2-3kb jump	Solexa-39450	492188542	49711042742	50
2-3kb jump	Solexa-39462	217999666	22017966266	22
5kb jump	Solexa-50902	147317752	14879092952	15
7kb jump	Solexa-50914	158260012	15984261212	16
9kb jump	Solexa-50937	143454662	14488920862	14
11kb jump	Solexa-51379	114671088	11581779888	12
fosmid jump	Solexa-46074	38364464	2762241408	2.8
	Total	1909866518	191783948862	192

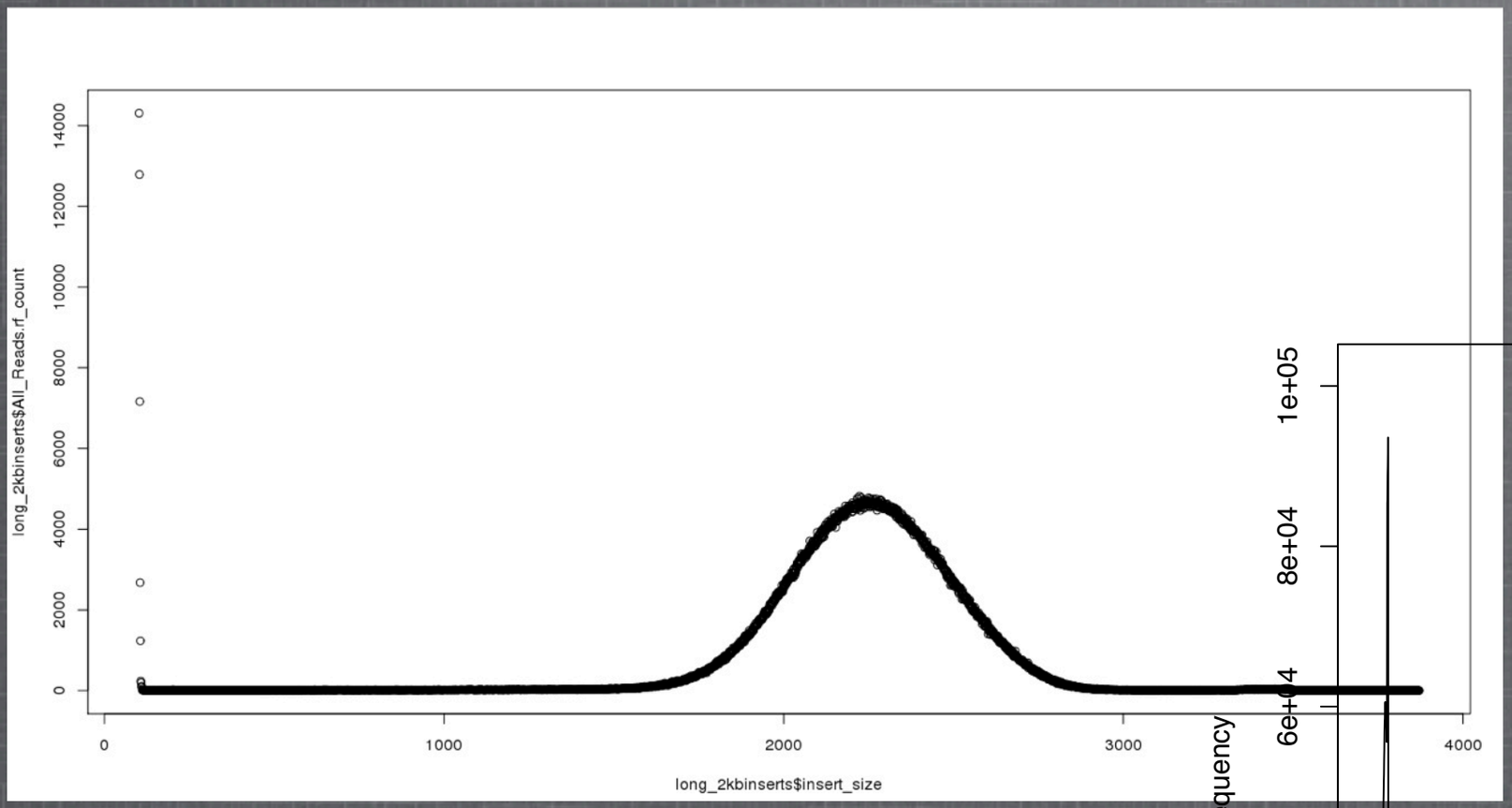
2-3KB FRAGMENT SIZES W/ BOWTIE2 MAPPING

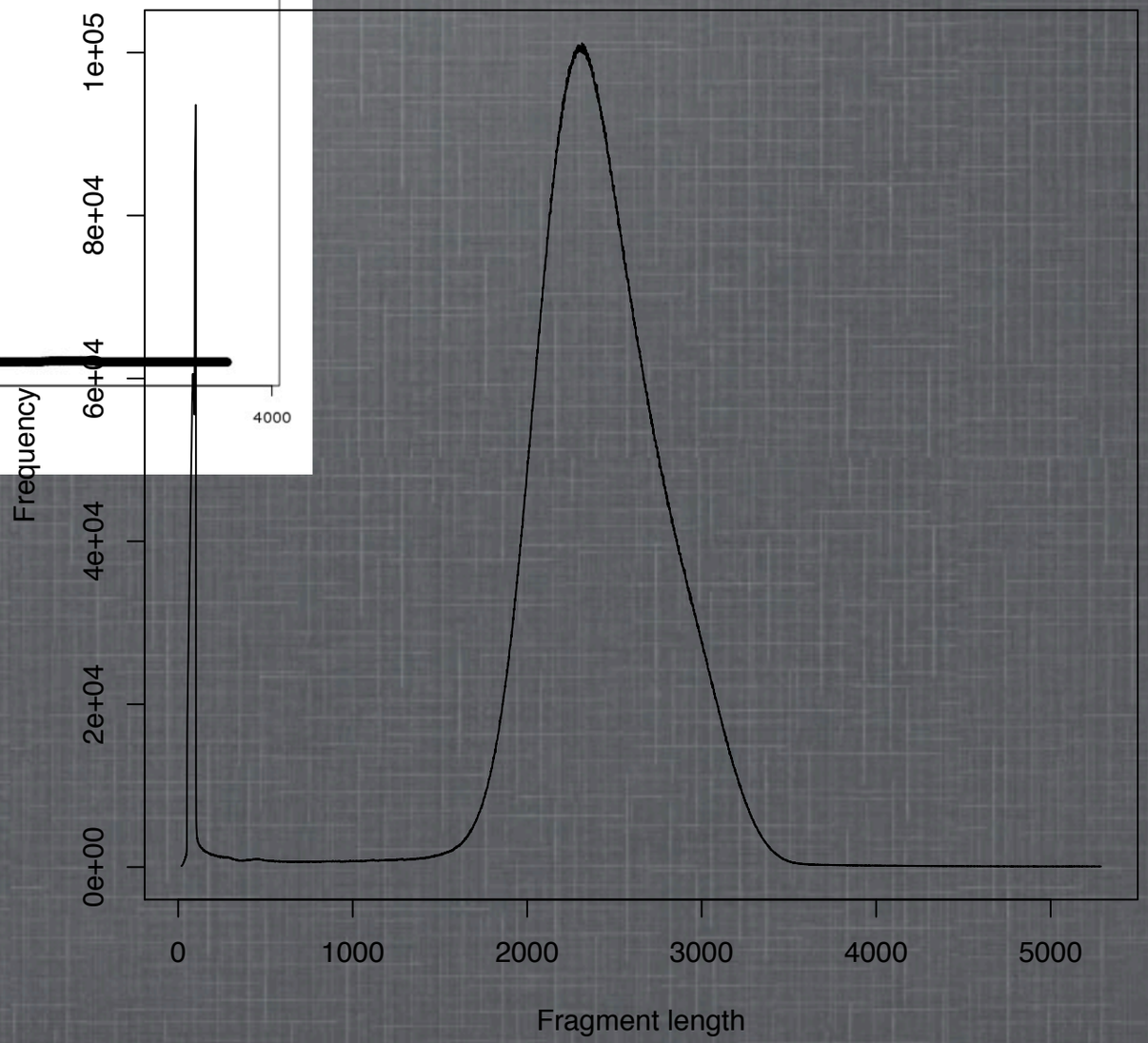


SMALTMAP

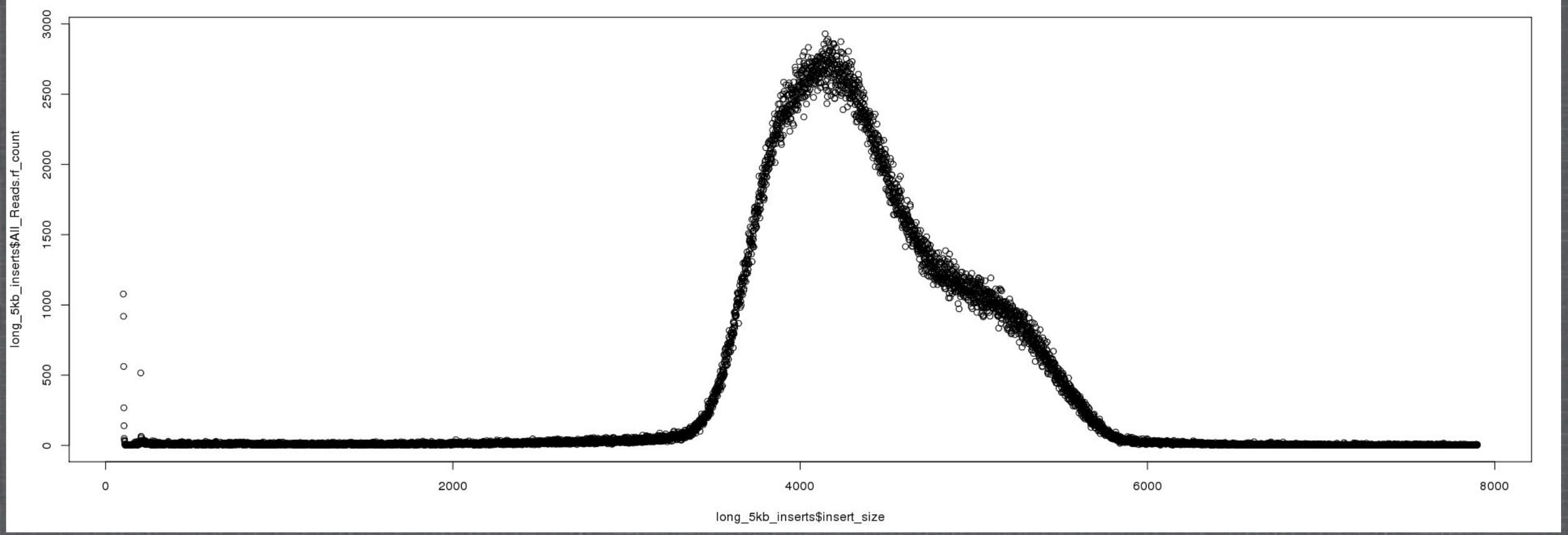
From the **REAPR** manual:
"We recommend SMALT (http://www.sanger.ac.uk/resources/software/smalt/)with the options -x -r. This will map reads repetitively (-r) and map each reads in a pair independently of each other (-x). Independent mapping is important, so that reads in a pair are not incorrectly forced to be mapped near to each other."

2-3KB
BOWTIE2 VS SMALTMAP

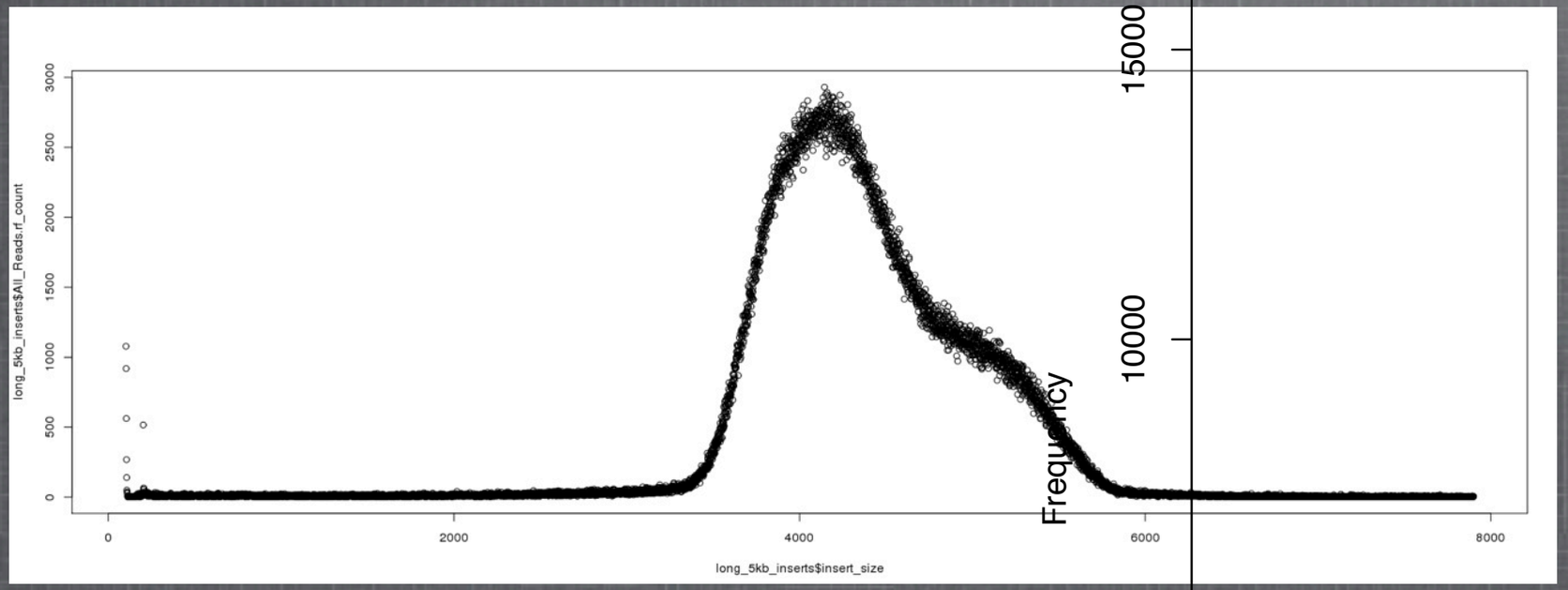


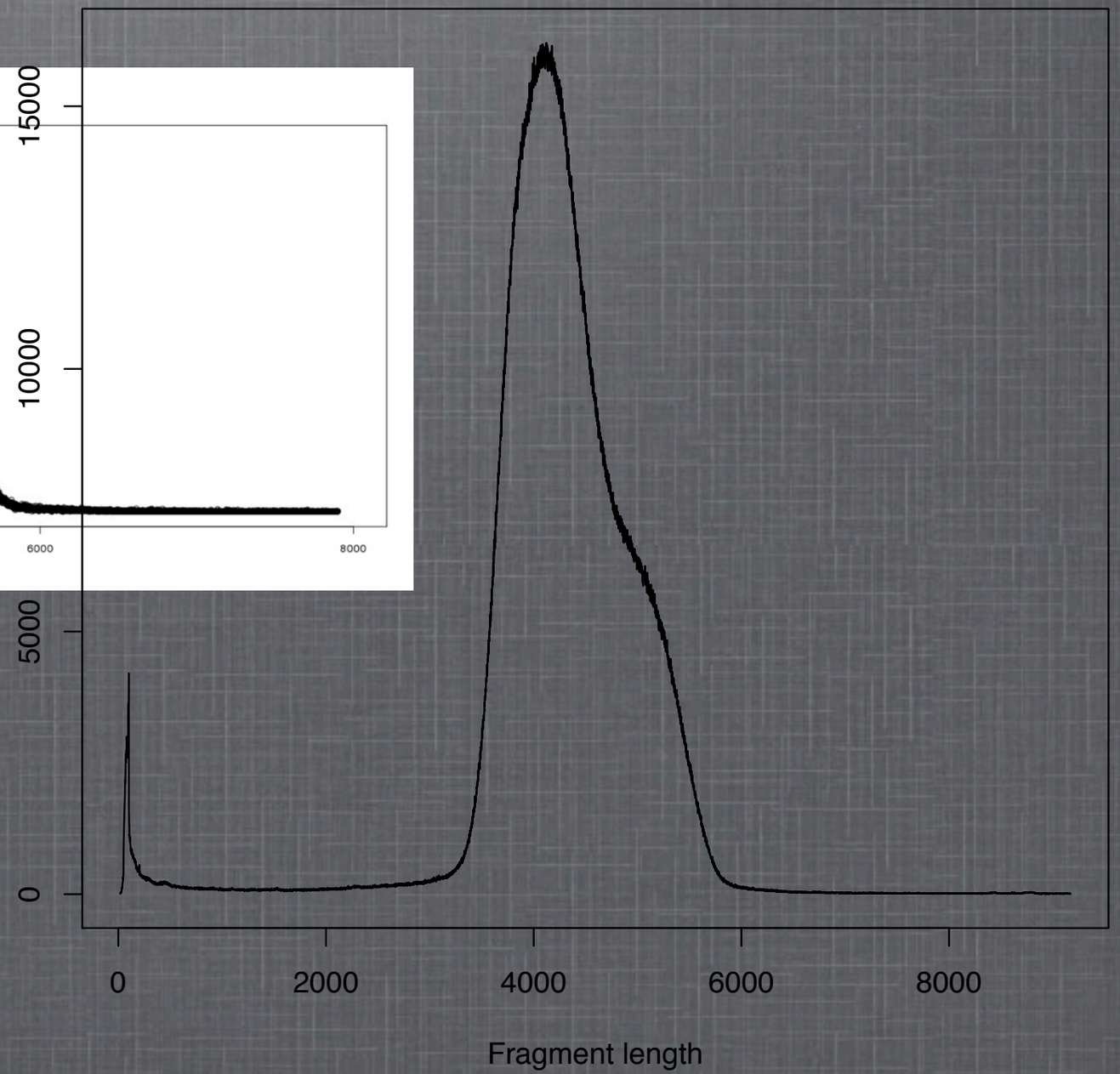


5KB FRAGMENT SIZES W/ BOWTIE2 MAPPING

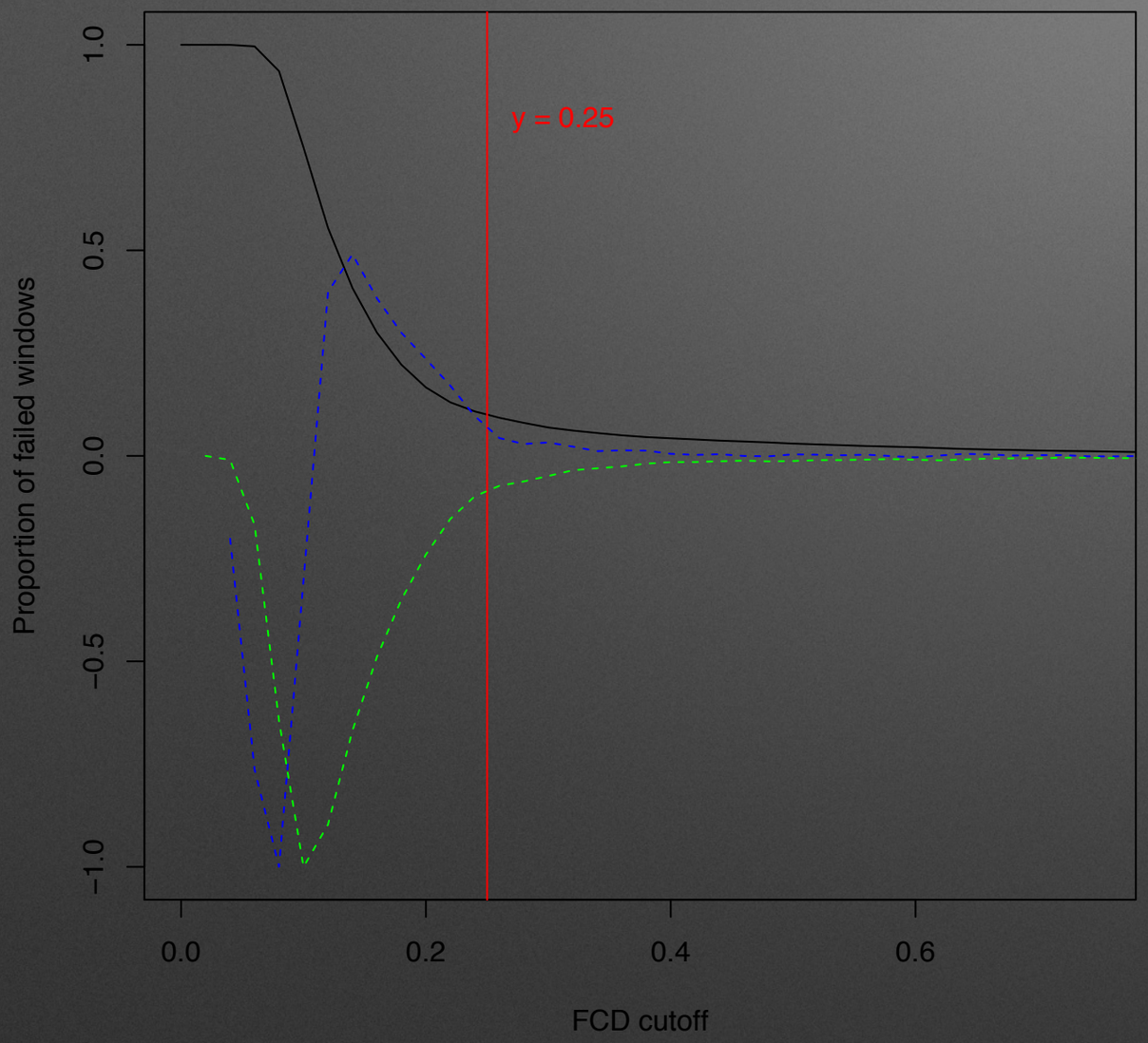


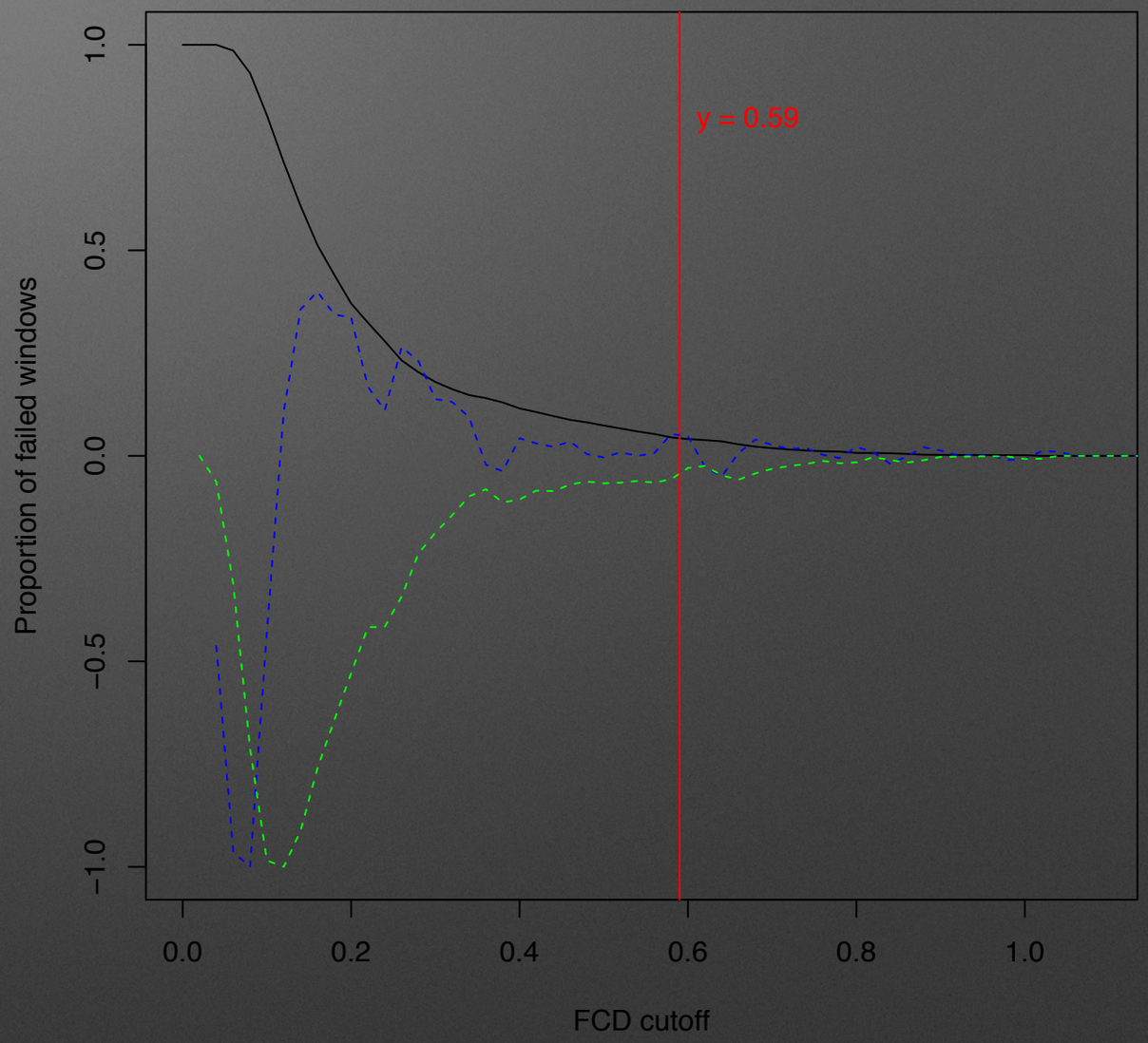
5KB
BOWTIE2 VS SMALTMAP





Some results..



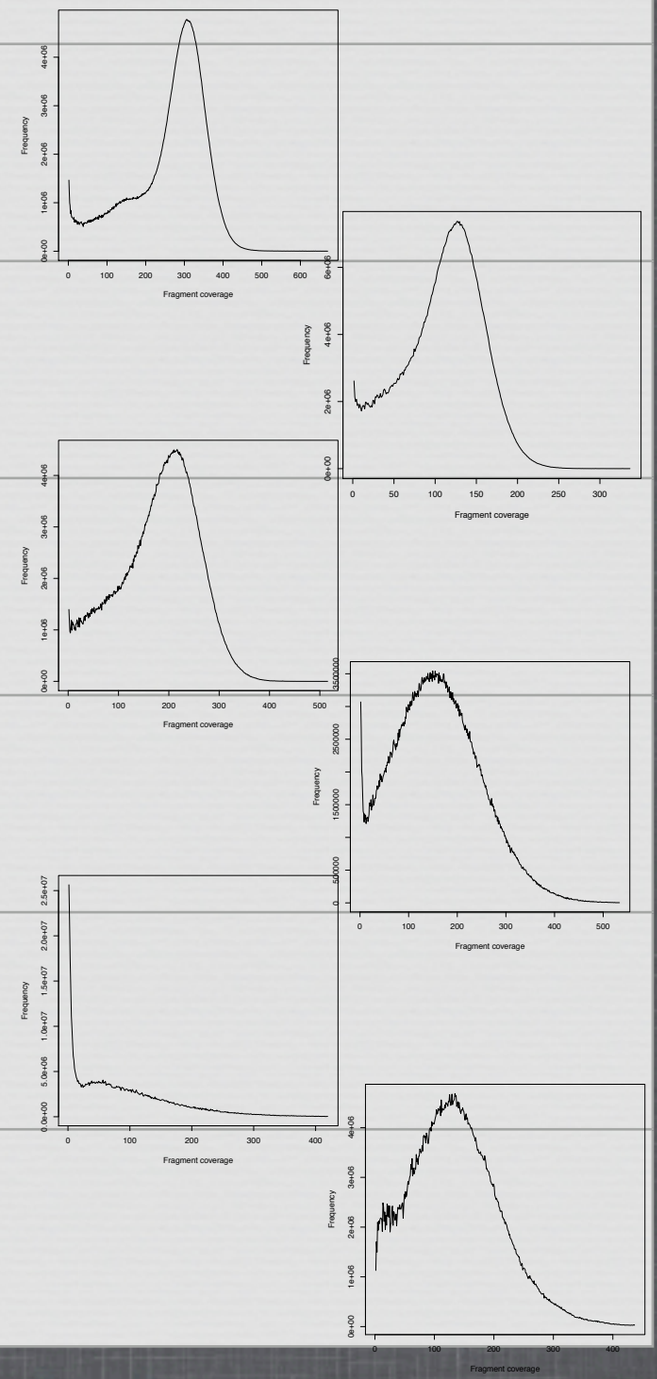


- 2-3kb library

40kb library

REAPR BREAKS

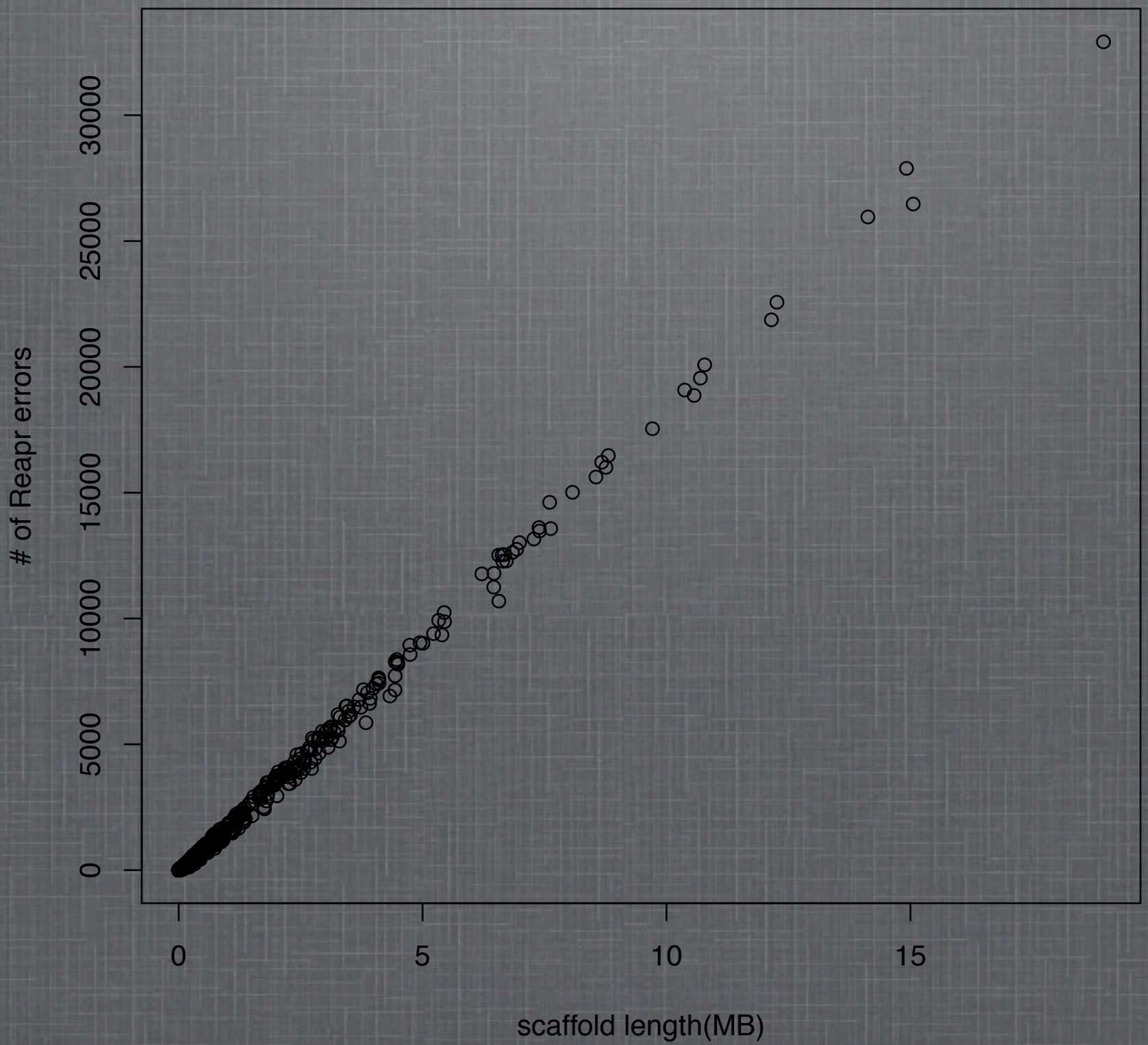
Library	Number of breaks	Fragment coverage
2-3kb	10125	~300
5kb	11064	~125
7kb	10104	~210
9kb	5810	~150
11kb	21317	~50?
40kb	4496	~150



REAPR BREAKS

Assembly	Number of scaffolds	N50	Total length	#gaps
Broad	-	3699709	848776495	68336
Reapr 2-3kb	10125	332860	846824676	63432
Reapr 5kb	11064	352213	841274373	61552
Reapr 7kb	10104	419373	840481238	62207
Reapr 9kb	5810	1482455	840875156	66440
Reapr 11kb	21317	98122	816288263	57532
Reapr 40kb	4496	2592500	847453497	67597

NUMBER OF ERRORS PER SCAFFOLD (40KB BREAK)



REAPR BREAKS IN COMMON

	Reapr 2-3kb	Reapr 5kb	Reapr 7kb	Reapr 9kb	Reapr 11kb	Reapr 40kb
Reapr 2-3kb	-					
Reapr 5kb	3260	-				
Reapr 7kb	2692	3658	-			
Reapr 9kb	1715	1856	1965	-		
Reapr 11kb	2703	3224	3094	2142	-	
Reapr 40kb	1183	1201	1215	1200	1233	-

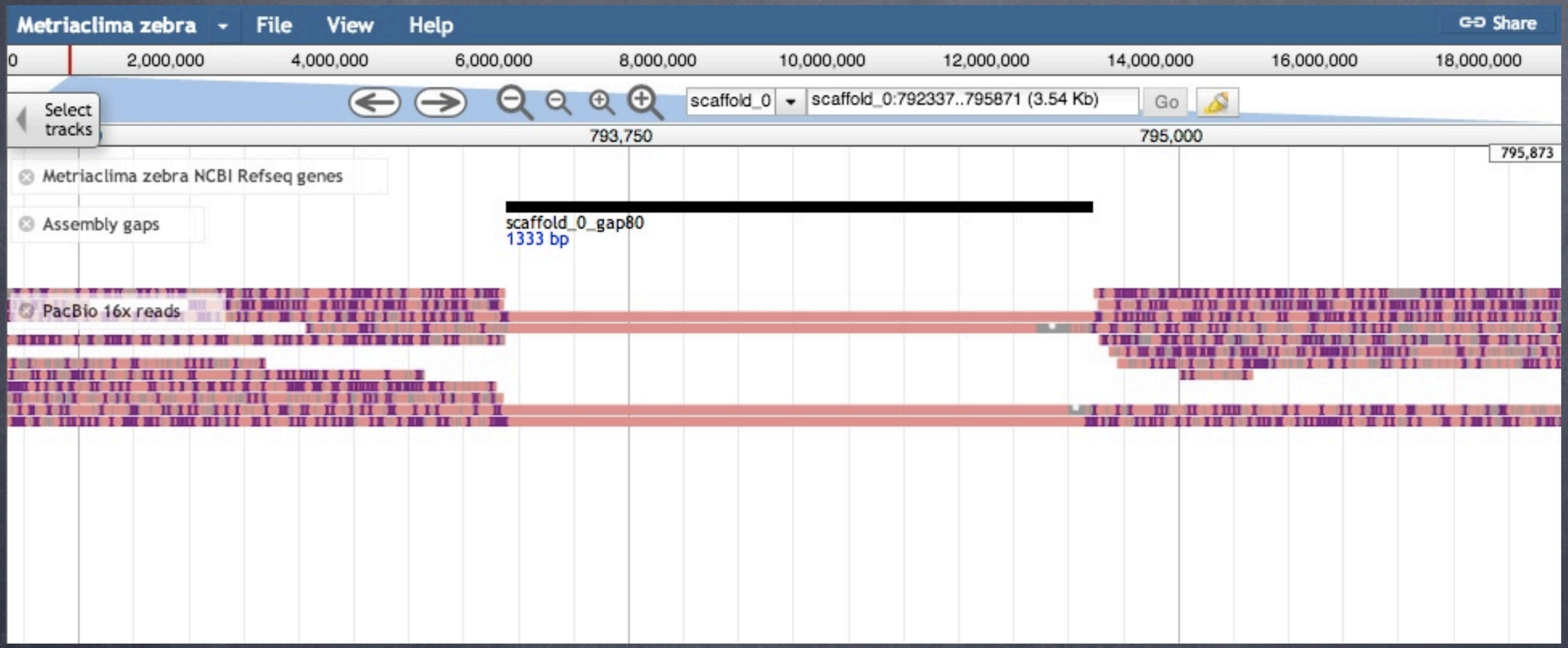
ASSEMBLY COMPARISON

Assembly	Broad			
Number of scaffolds	3,750			
Total size of scaffolds	848,776,495			
Total scaffold length as % of assumed genome size	84.9			
Mean scaffold size	226,340			
N50 scaffold length	3,699,709			
N50 scaffold - NG50 scaffold length difference	692,019			
scaffold %N	15.93			
Number of gaps	68,336			

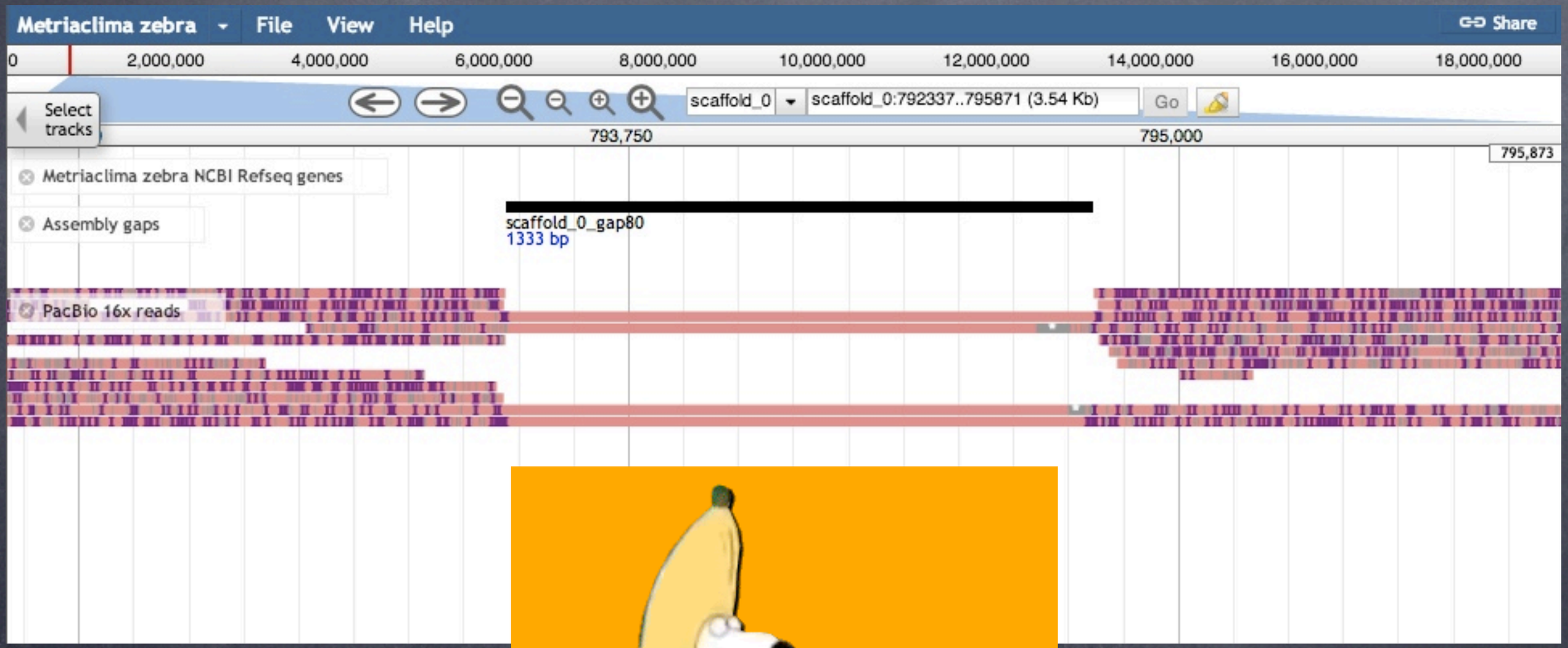
ASSEMBLY COMPARISON

Assembly	Broad	40kb REAPR broken		
Number of scaffolds	**3,750**	4,496		
Total size of scaffolds	**848,776,495**	847,453,497		
Total scaffold length as % of assumed genome size	**84.9**	84.7		
Mean scaffold size	**226,340**	188,491		
N50 scaffold length	**3,699,709**	2,592,500		
N50 scaffold - NG50 scaffold length difference	**692,019**	343,997		
scaffold %N	15.93	**15.8**		
Number of gaps	68,336	**67,597**		

PBJelly


Metriaclima zebra
File
View
Help
Share
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
12,000,000
14,000,000
16,000,000
18,000,000
Select tracks
scaffold_0
scaffold_0:792337..795871 (3.54 Kb)
Go
793,750
795,000
795,873
Metriaclima zebra NCBI Refseq genes
Assembly gaps
scaffold_0_gap80
1333 bp
PacBio 16x reads

PBJelly


Metriaclima zebra
File
View
Help
Share
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
12,000,000
14,000,000
16,000,000
18,000,000
Select tracks
scaffold_0
scaffold_0:792337..795871 (3.54 Kb)
Go
793,750
795,000
795,873
Metriaclima zebra NCBI Refseq genes
Assembly gaps
scaffold_0_gap80
1333 bp
PacBio 16x reads


IT'S PEANUT BUTTER JELLY TIME!!!

PBJelly

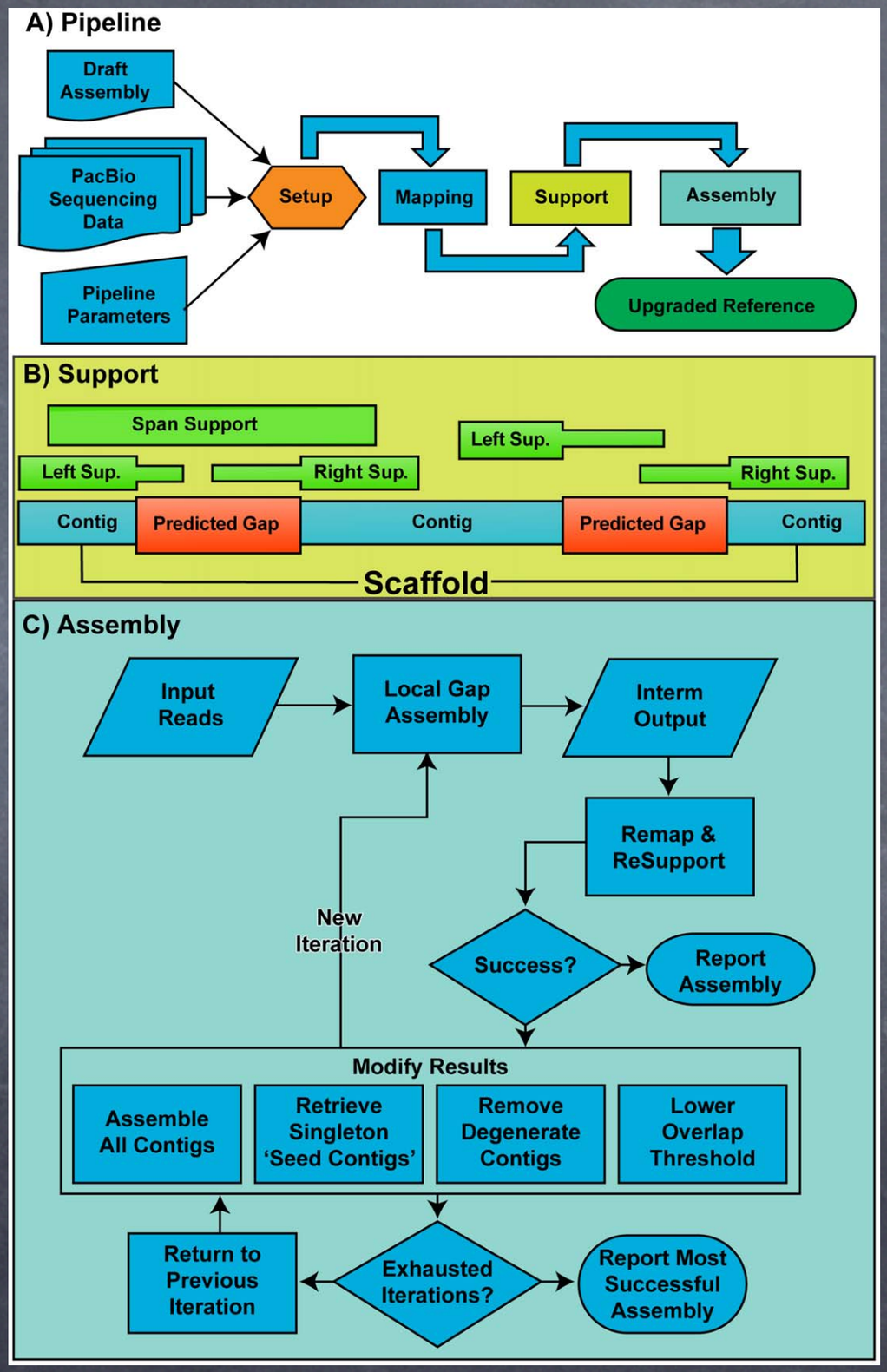


Figure 1. A schematic of PBJelly's workflow and decision-making. (A) A flow chart of PBJelly's steps. (B) A schematic describing two hypothetical gaps supported by reads and the classifications used during the Support step. (C) A detailed flow chart for local assembly of PacBio reads in a gap region used during the assembly step.
doi:10.1371/journal.pone.0047768.g001

ASSEMBLY COMPARISON

Assembly	Broad	40kb REAPR broken	40kb REAPR+PBJelly	
Number of scaffolds	3,750	4,496	**3,531**	
Total size of scaffolds	848,776,495	847,453,497	**868,183,531**	
Total scaffold length as % of assumed genome size	84.9	84.7	**86.8**	
Mean scaffold size	226,340	188,491	**245,875**	
N50 scaffold length	**3,699,709**	2,592,500	3,314,734	
N50 scaffold - NG50 scaffold length difference	**692,019**	343,997	559,051	
scaffold %N	15.93	15.8	**4.65**	
Number of gaps	68,336	67,597	**12,030**	

ASSEMBLY COMPARISON

Assembly	Broad	40kb REAPR broken	40kb REAPR+PBJelly	PBJelly on Broad
Number of scaffolds	3,750	4,496	3,531	**3,265**
Total size of scaffolds	848,776,495	847,453,497	**868,183,531**	861,582,306
Total scaffold length as % of assumed genome size	84.9	84.7	**86.8**	86.2
Mean scaffold size	226,340	188,491	245,875	**263,884**
N50 scaffold length	3,699,709	2,592,500	3,314,734	**4,092,986**
N50 scaffold - NG50 scaffold length difference	692,019	343,997	559,051	**3,284,891**
scaffold %N	15.93	15.8	**4.65**	4.9
Number of gaps	68,336	67,597	12,030	**11,230**

A PEAK AT LG7 (KNOWN MISASSEMBLY)

Reapr 2-3kb	>scaffold_0_3699936_4471845
Reapr 5kb	>scaffold_0_3699936_3718203
Reapr 7kb	>scaffold_0_3699936_3700656
Reapr 9kb	>scaffold_0_3699936_3700524
Reapr 11kb	>scaffold_0_3699936_3841628
Reapr 40kb	>scaffold_0_3699936_13877085

Broad_40kb_62F6HAAXX.1

Broad_2.5kb_801KYABXX.2

Broad_2.5kb_801KYABXX.3

Broad_2.5kb_801KYABXX.4

Broad_5kb_803DNABXX.6

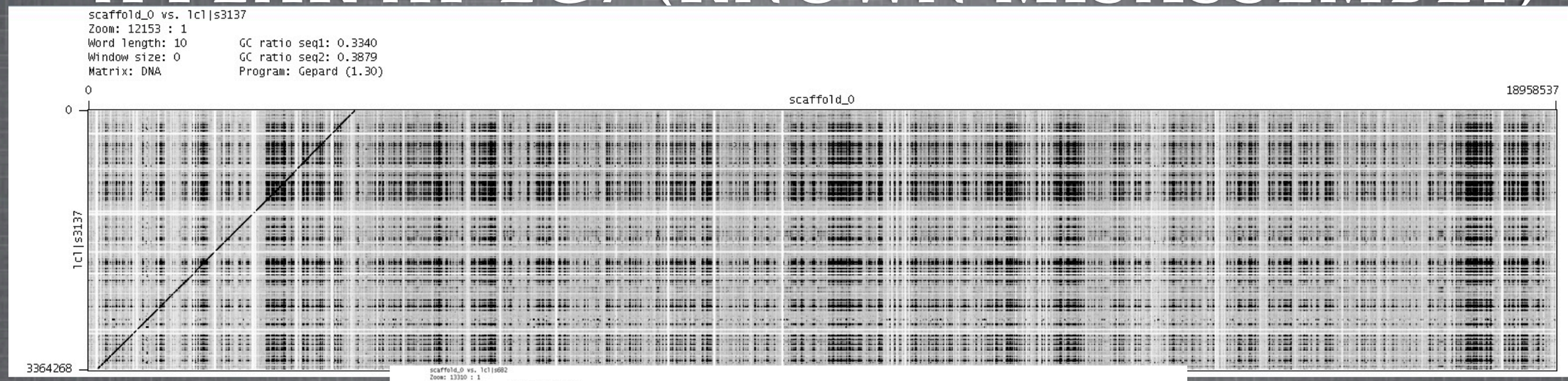
Broad_7kb_803DNABXX.1

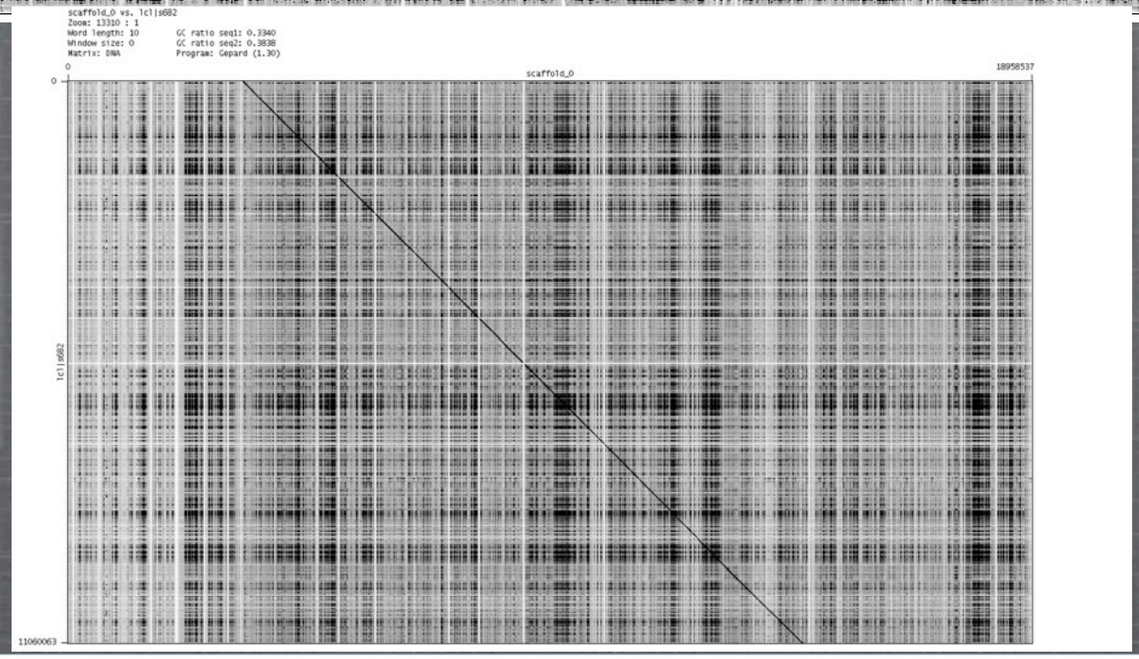
Broad_9kb_803DNABXX.2

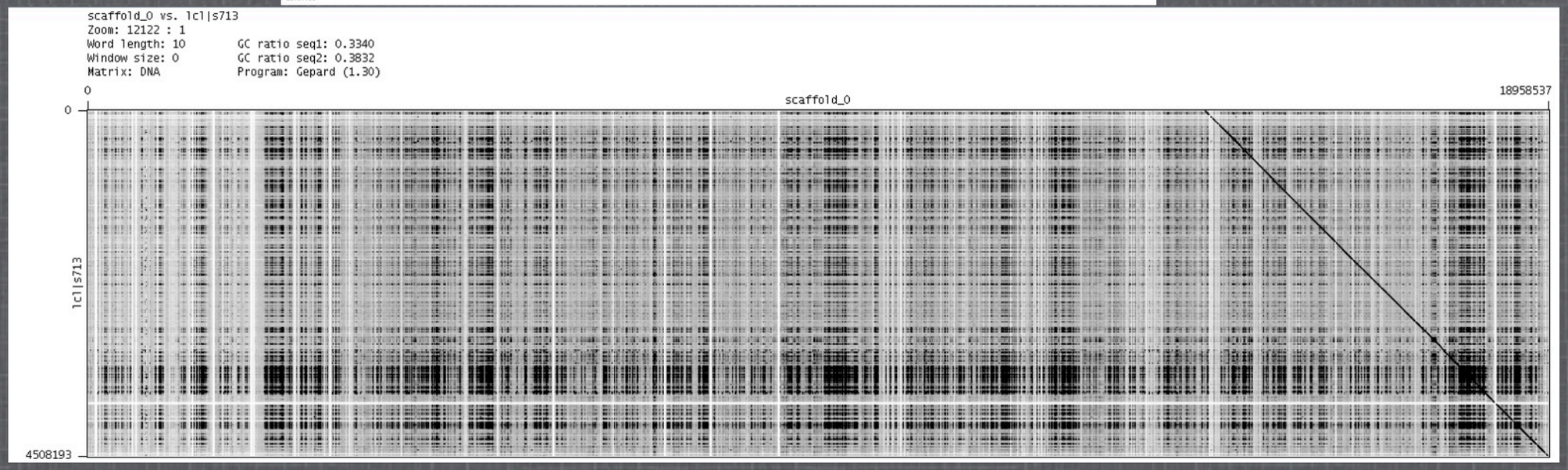
A PEAK AT LG7 (KNOWN MISASSEMBLY)

>scaffold_0_13878029_14447808_scaffold_0_3699936_13877085_scaffold_0_3436025_3699617
>scaffold_0_14479301_14480673_scaffold_0_14458627_14463568_scaffold_0_14467772_14475683
>scaffold_0_14482037_16479292_scaffold_0_16483848_18958539
>scaffold_0_3433764_3435224
>scaffold_0_14456193_14457333
>scaffold_0_3431160_3432957
>scaffold_0_660314_2394353_scaffold_0_2395738_3427604_scaffold_0_78142_648570_scaffold_0_342923
1_3430274_scaffold_0_656196_658695_scaffold_0_653230_655132_scaffold_0_75433_77771
>scaffold_0_62640_72384_scaffold_0_1_61220
>scaffold_0_649929_652193

A PEAK AT LG7 (KNOWN MISASSEMBLY)







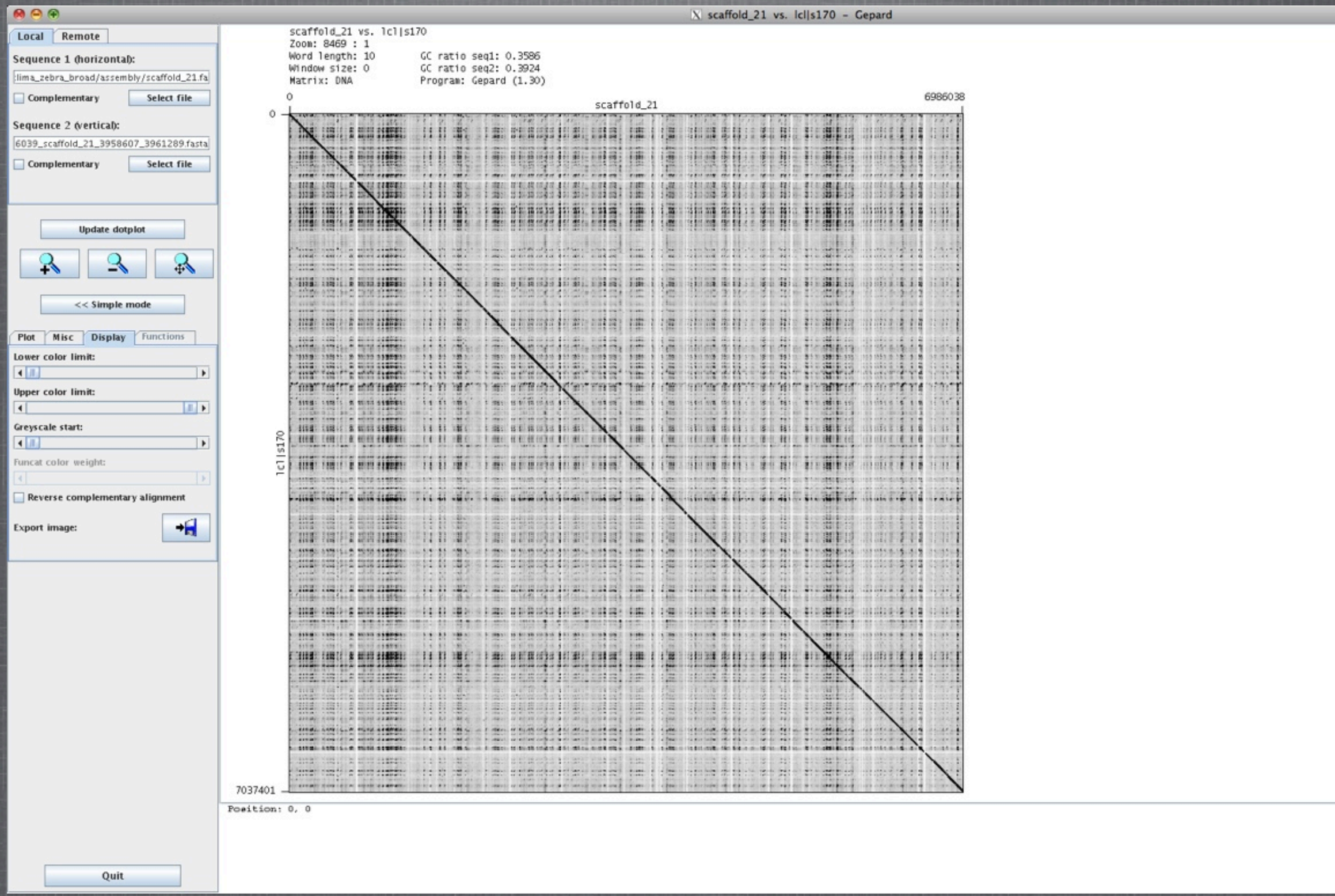
A PEAK AT LG7 SEX REGION

>scaffold_0_13878029_14447808_scaffold_0_3699936_13877085_scaffold_0_3436025_3699617
>scaffold_0_14479301_14480673_scaffold_0_14458627_14463568_scaffold_0_14467772_14475683
>scaffold_0_14482037_16479292_scaffold_0_16483848_18958539
>scaffold_0_3433764_3435224
>scaffold_0_14456193_14457333
>scaffold_0_3431160_3432957
>scaffold_0_660314_2394353_scaffold_0_2395738_3427604_scaffold_0_78142_648570_scaffold_0_342923
1_3430274_scaffold_0_656196_658695_scaffold_0_653230_655132_scaffold_0_75433_77771
>scaffold_0_62640_72384_scaffold_0_1_61220
>scaffold_0_649929_652193

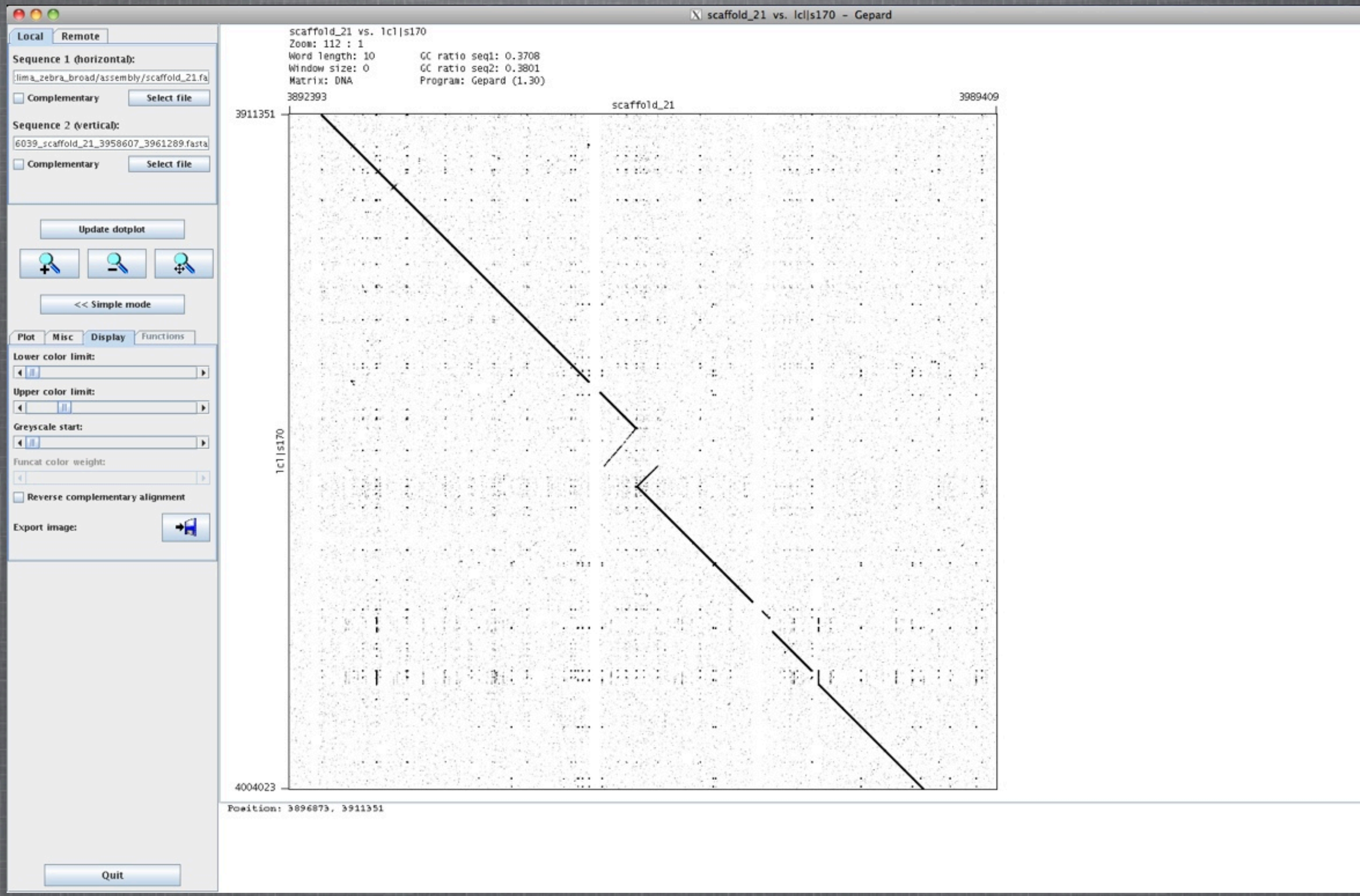
>scaffold_21_3957163_3958239
>scaffold_21_1_3955908_scaffold_21_3961310_3964005_scaffold_21_3964860_6986039_scaffold_21_395
8607_3961289

>scaffold_415_scaffold_1819_scaffold_2171_scaffold_3436_scaffold_2311

A PEAK AT LG7 SEX REGION



A PEAK AT LG7 SEX REGION



TODO

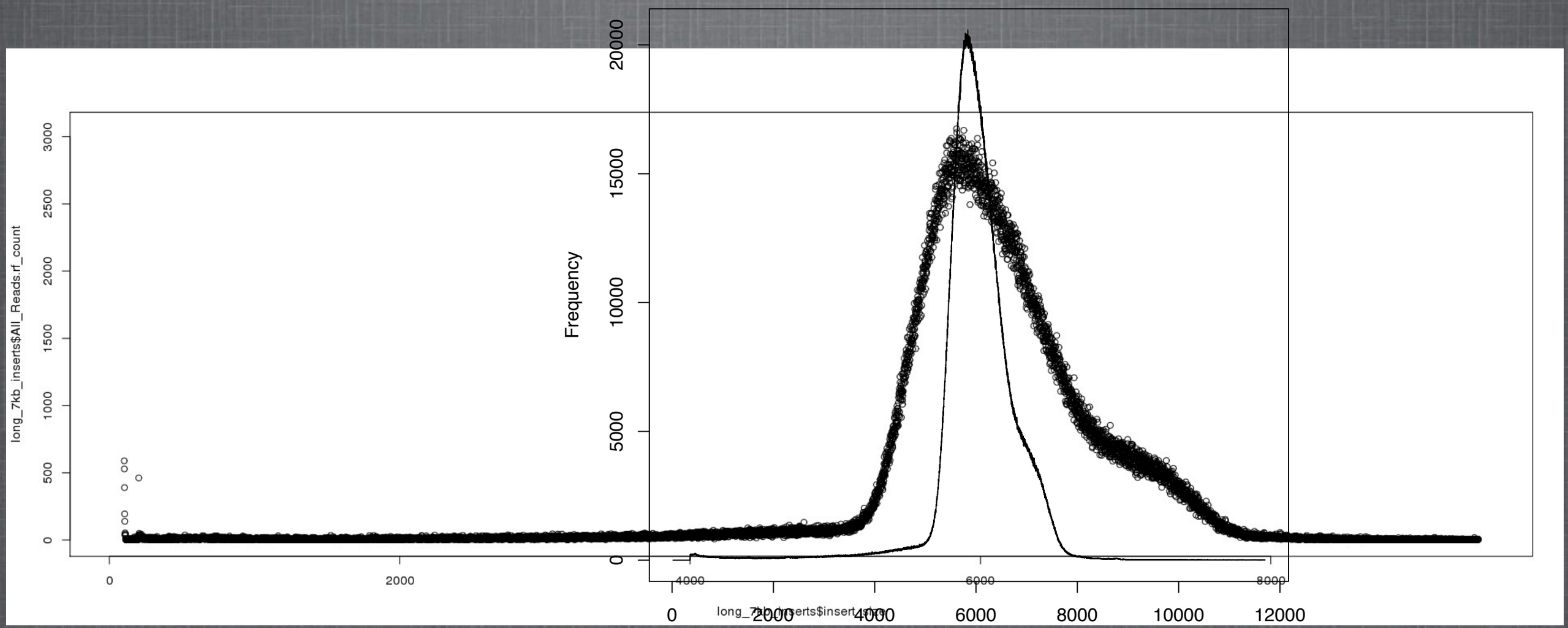
-Get “perfectmap” working, and rerun REAPR with it.
-Run PBJelly on other REAPR broken assemblies (est. 1-2 weeks)
-Reapr on PBJelly assembly
-is scaffold 21 rearrangement heterozygous in the male (pacbio male?)
-slope of FCD curves?
-Choose which broken + PBJelly assembly is most accurate
    -CEGMA analyses
    -RepeatMasker analyses
    -ALE?
    -# of stop codons removed
    -non syn subs (SNPEff things)

-Annotate with MAKER (Broad vs PBJelly, total number of exons?)
-LG7XY contigs?
-What else?
-Continue writing manuscript

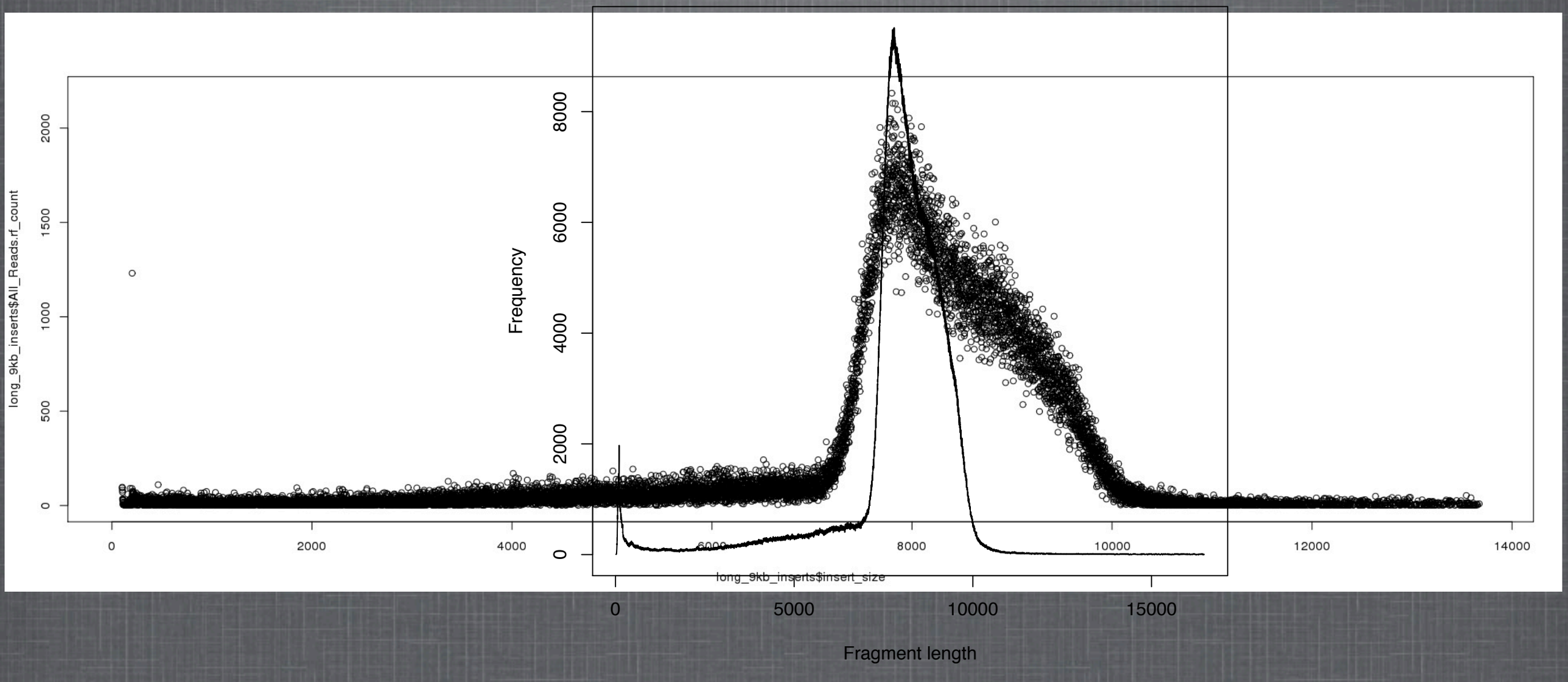
Time permitting... a little Blue Oyster Cult

- http://www.seehearparty.com/#g=computer&g=reaper&g=programming&g=science&s=https%3A%2F%2Fsoundcloud.com%2Fconsumable_electronica%2Fdont-fear-the-reaper-dancin (courtesy of Dr. Reade Roberts)

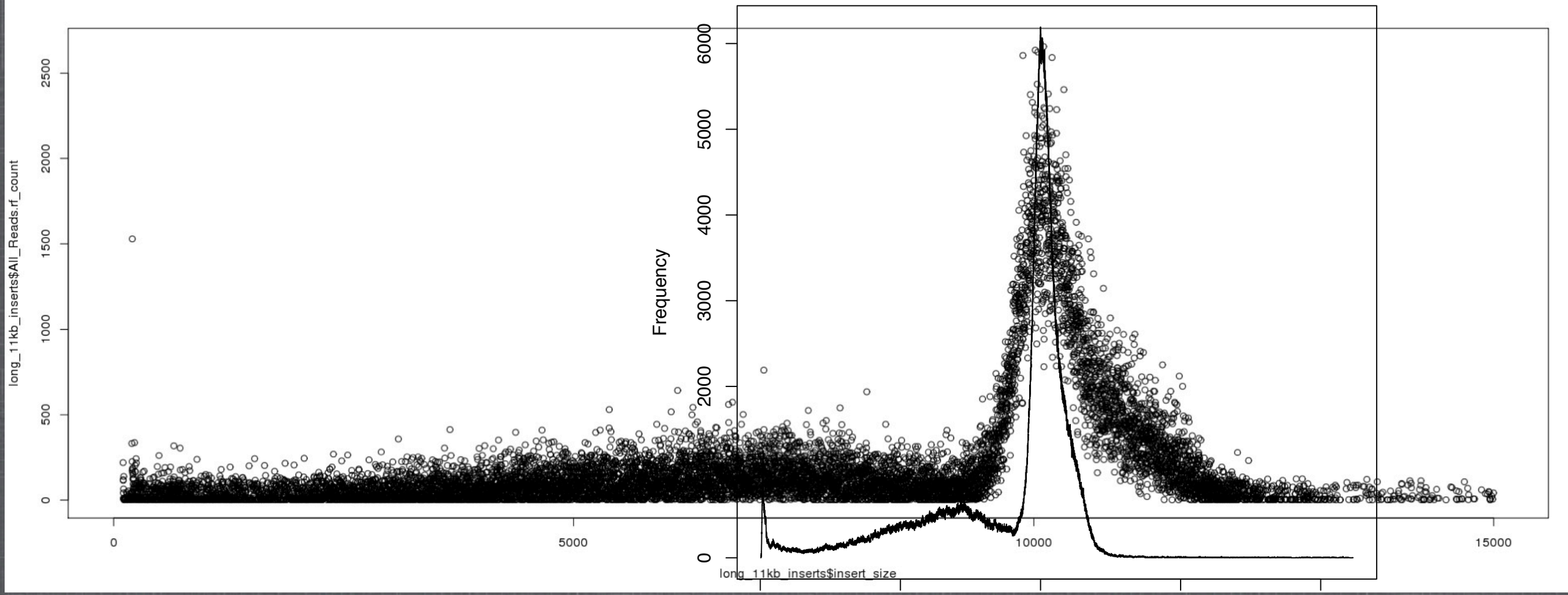
7KB FRAGMENT LENGTHS



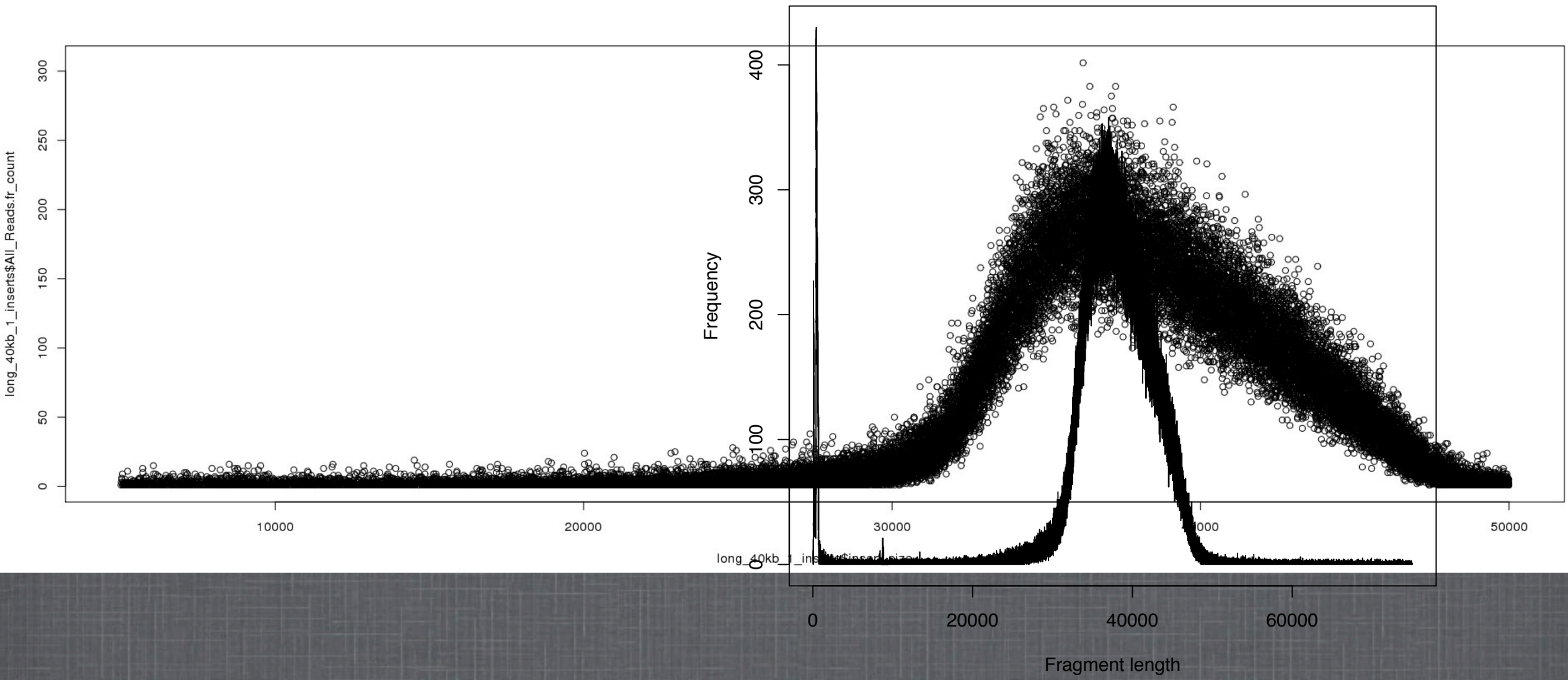
9KB FRAGMENT LENGTHS



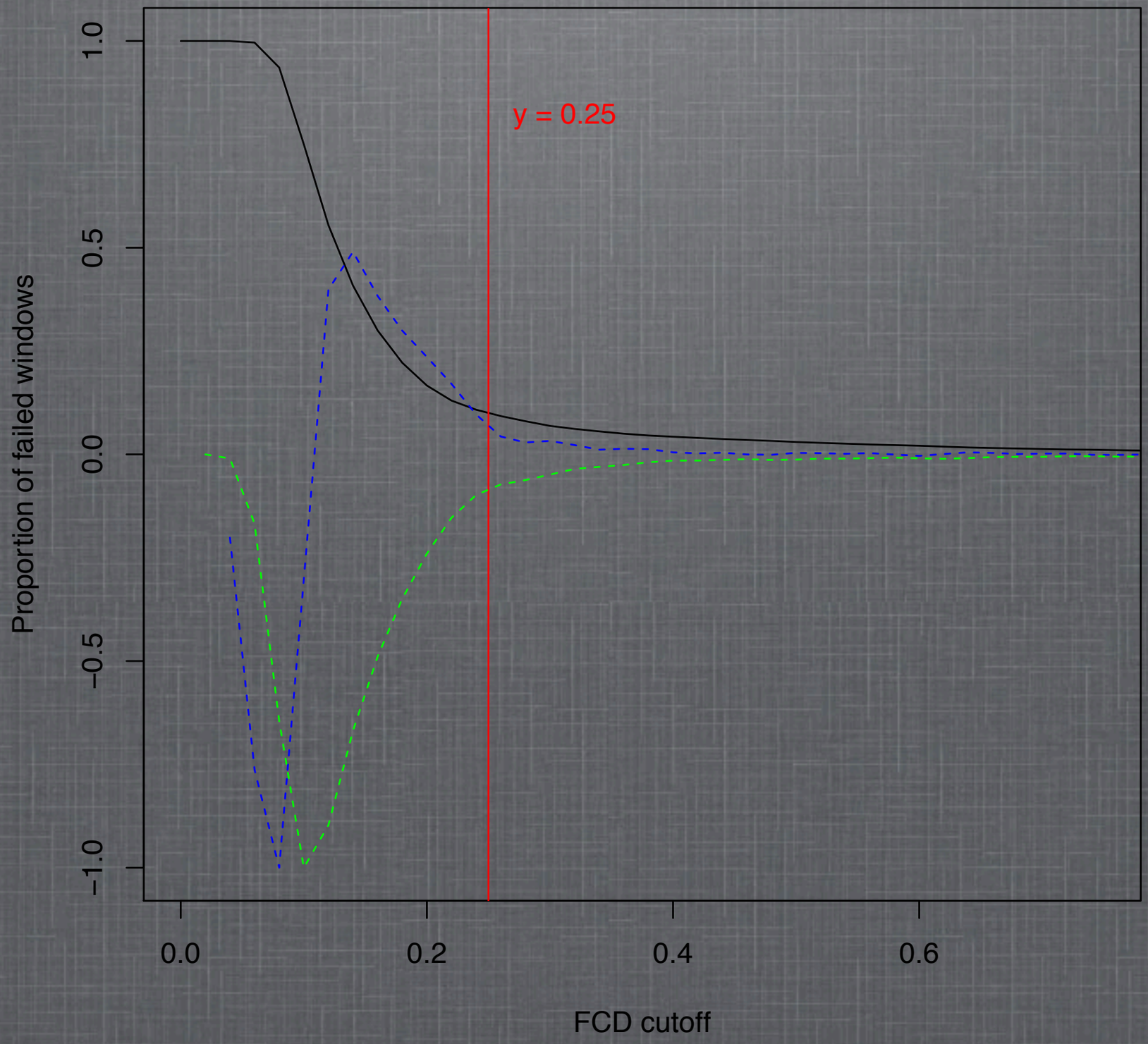
11KB FRAGMENT LENGTHS



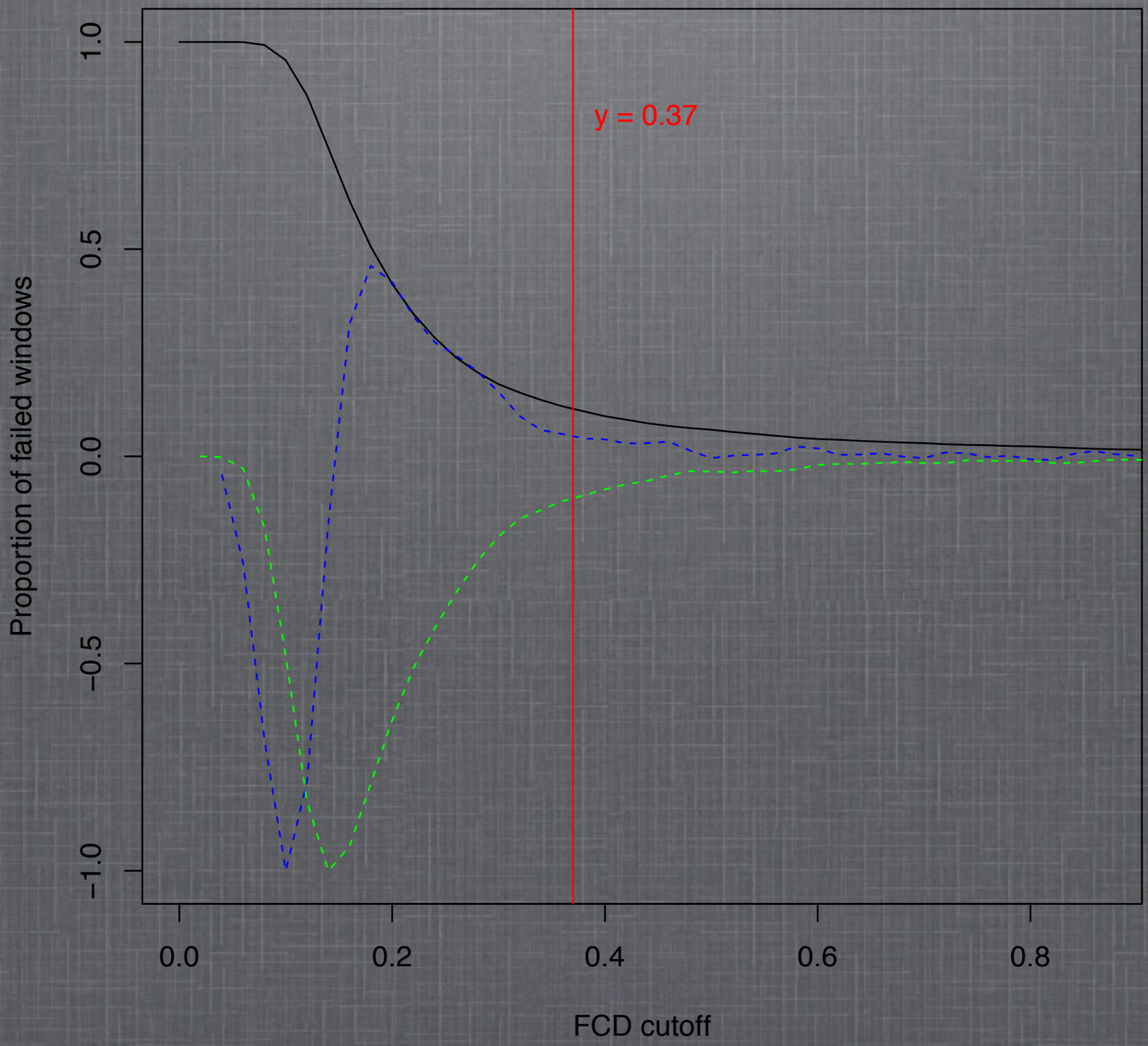
40KB FRAGMENT LENGTHS



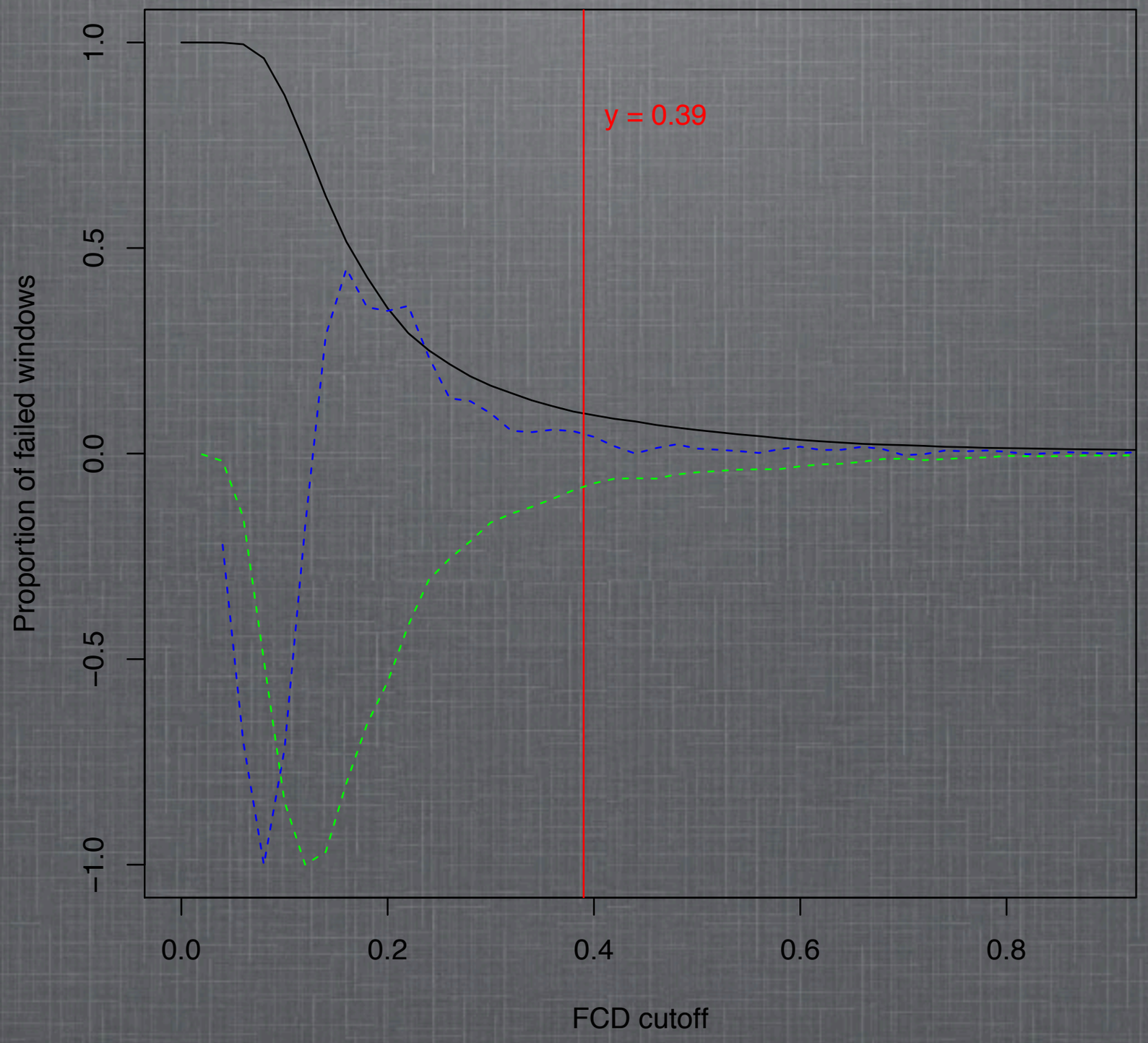
2-3KB FCD CUTOFF



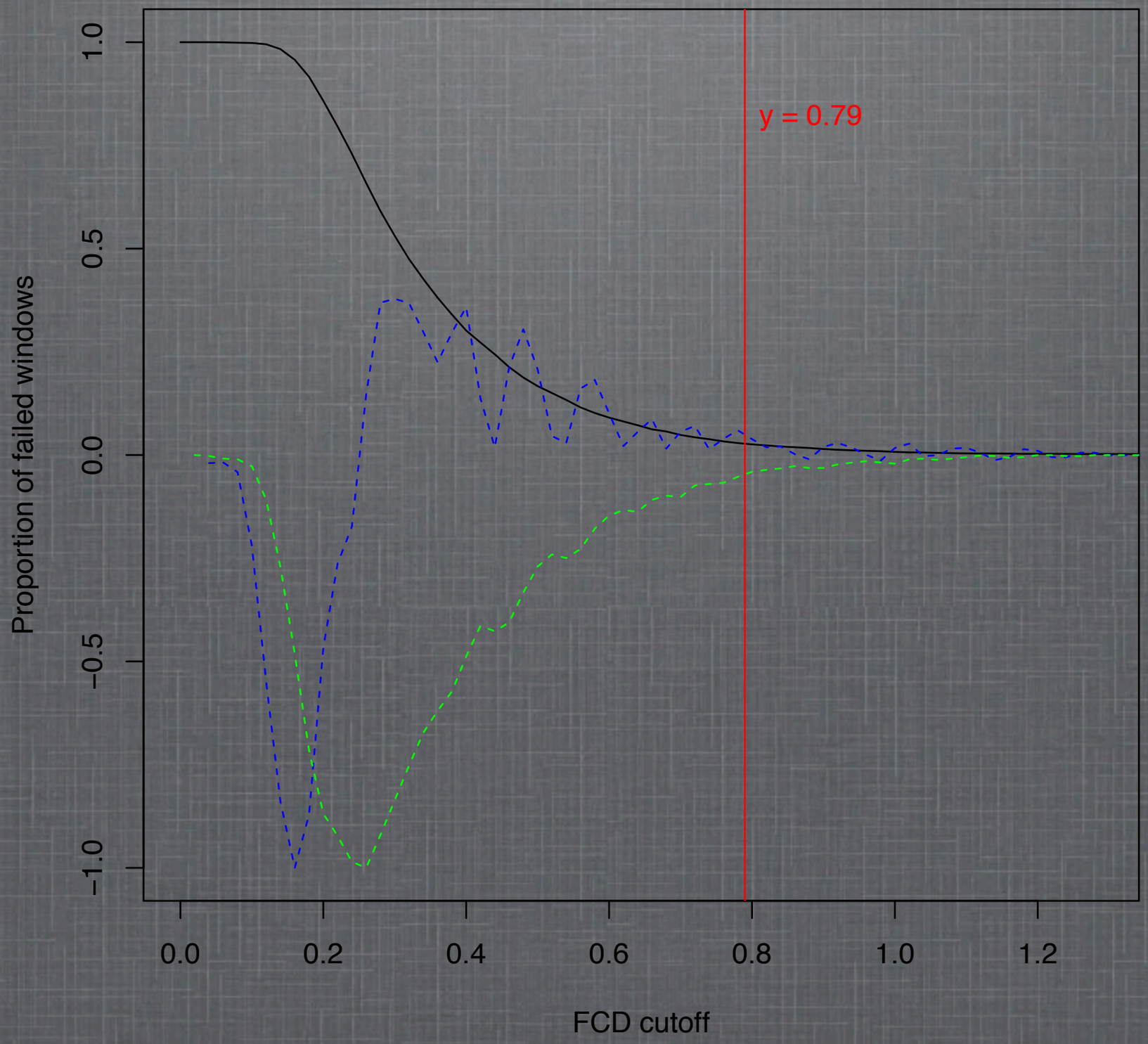
5KB FCD CUTOFF



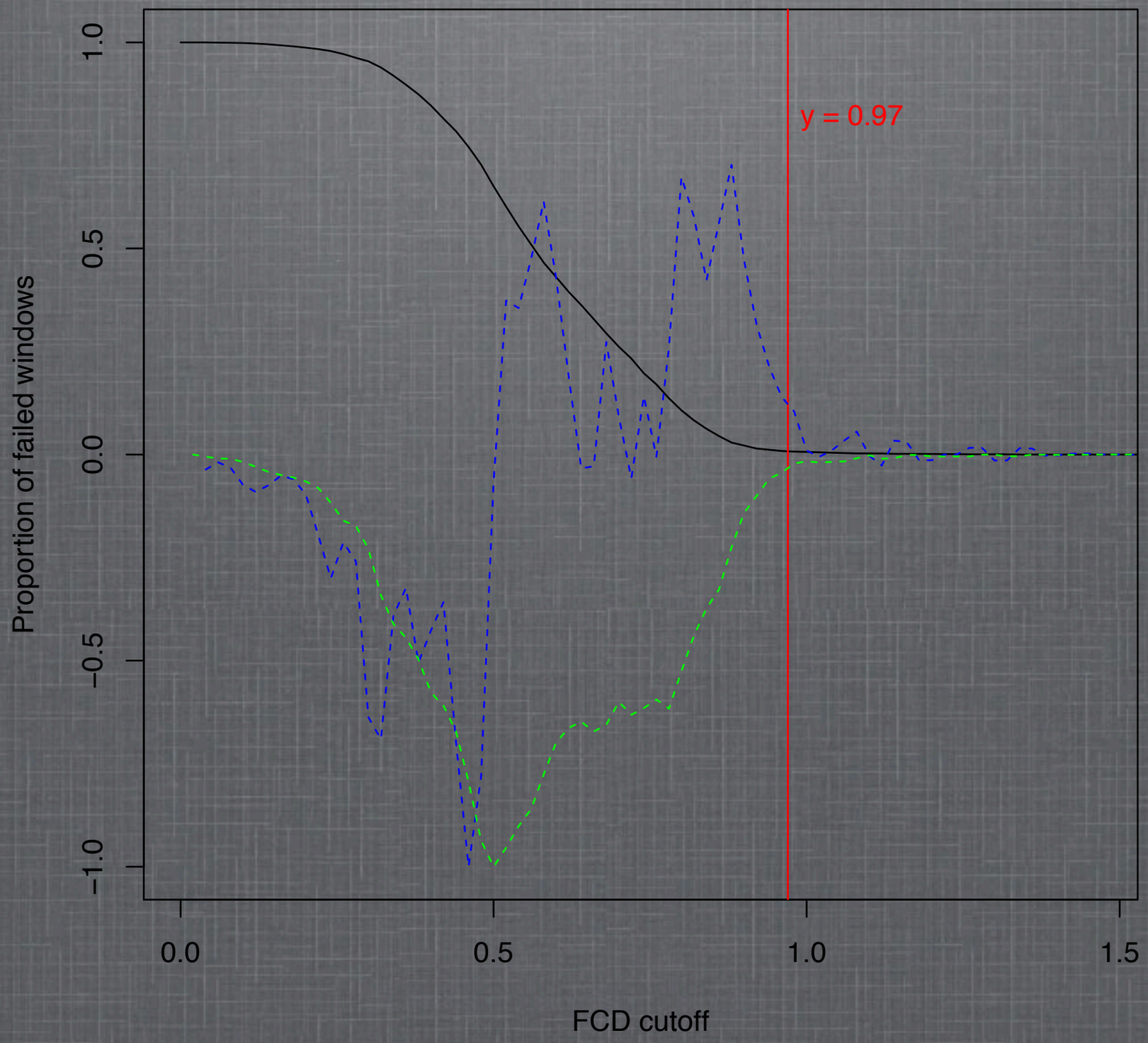
7KB FCD CUTOFF



9KB FCD CUTOFF



11KB FCD CUTOFF



40KB FCD CUTOFF