ScaffHiC – Genome Scaffolding by Modelling Distributions of HiC Paired End Reads

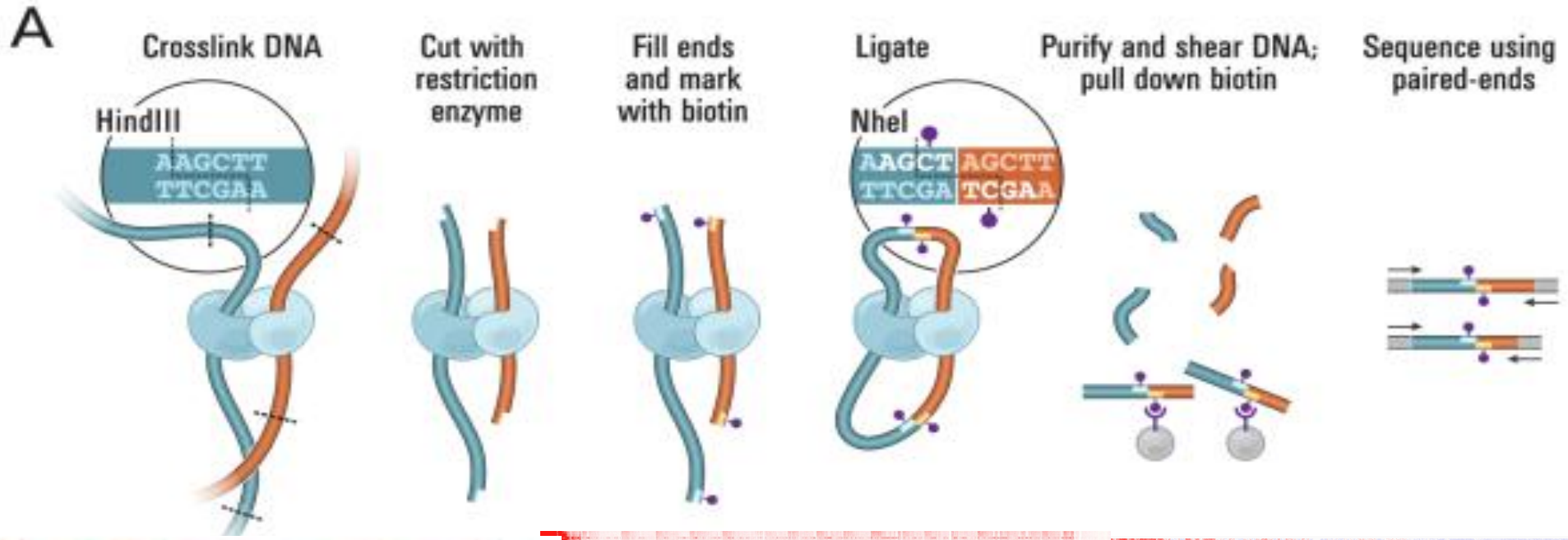
Zemin Ning

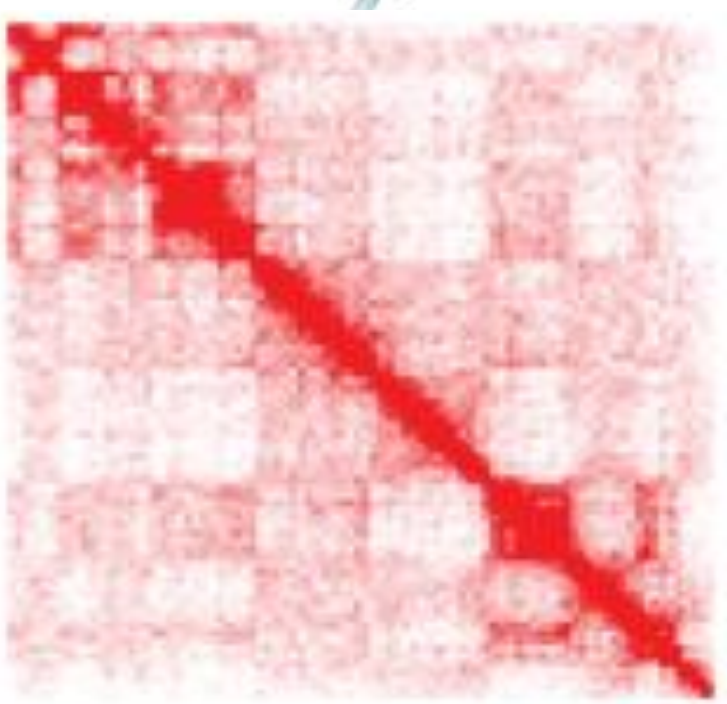
The Wellcome Sanger Institute

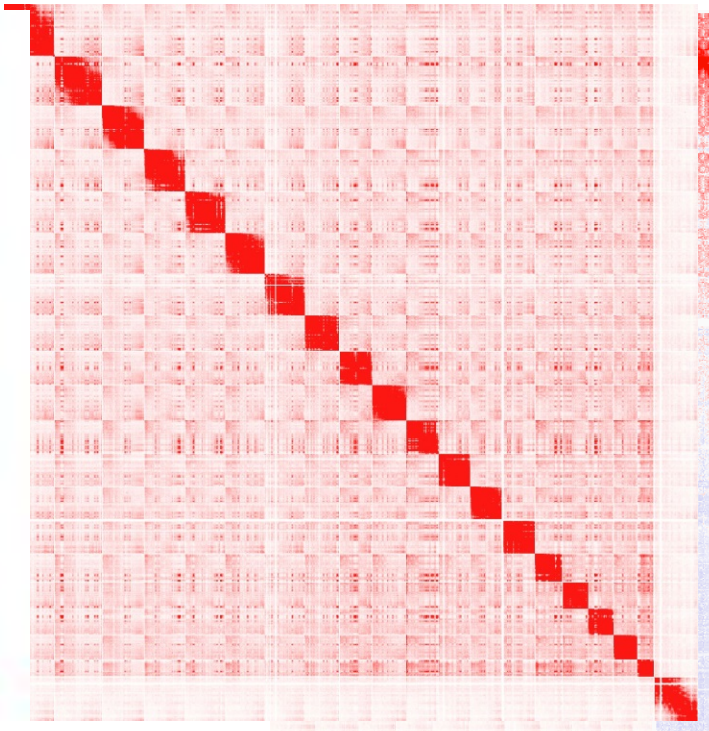
UK

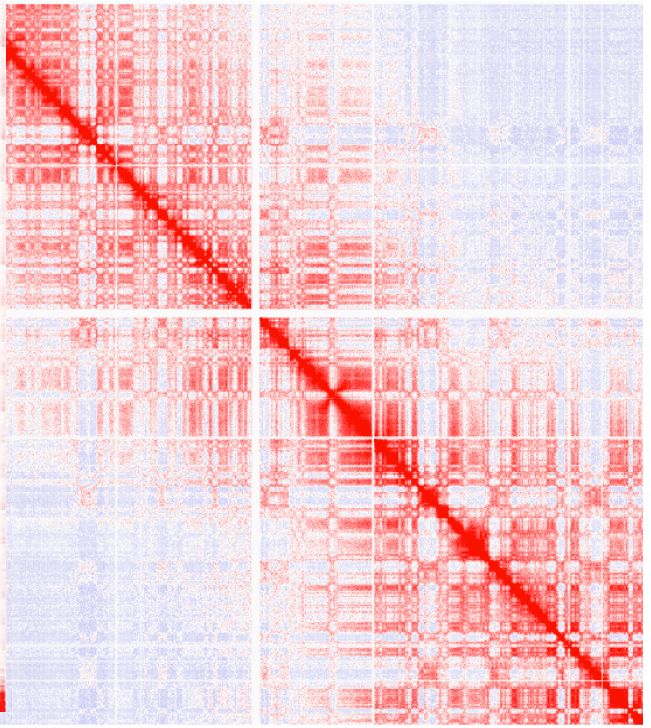
wellcome trust sanger institute

HiC – 3D Genomics

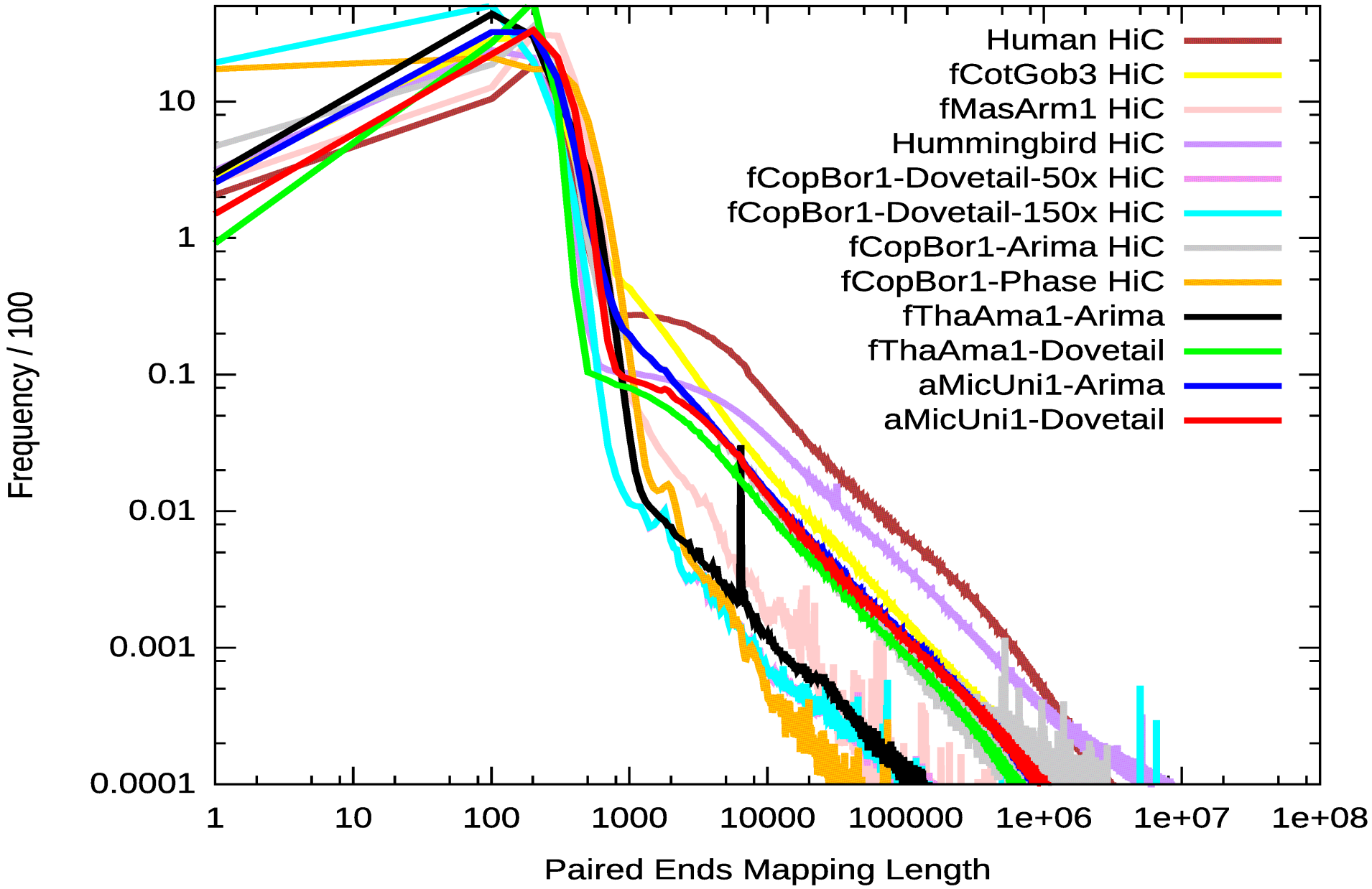




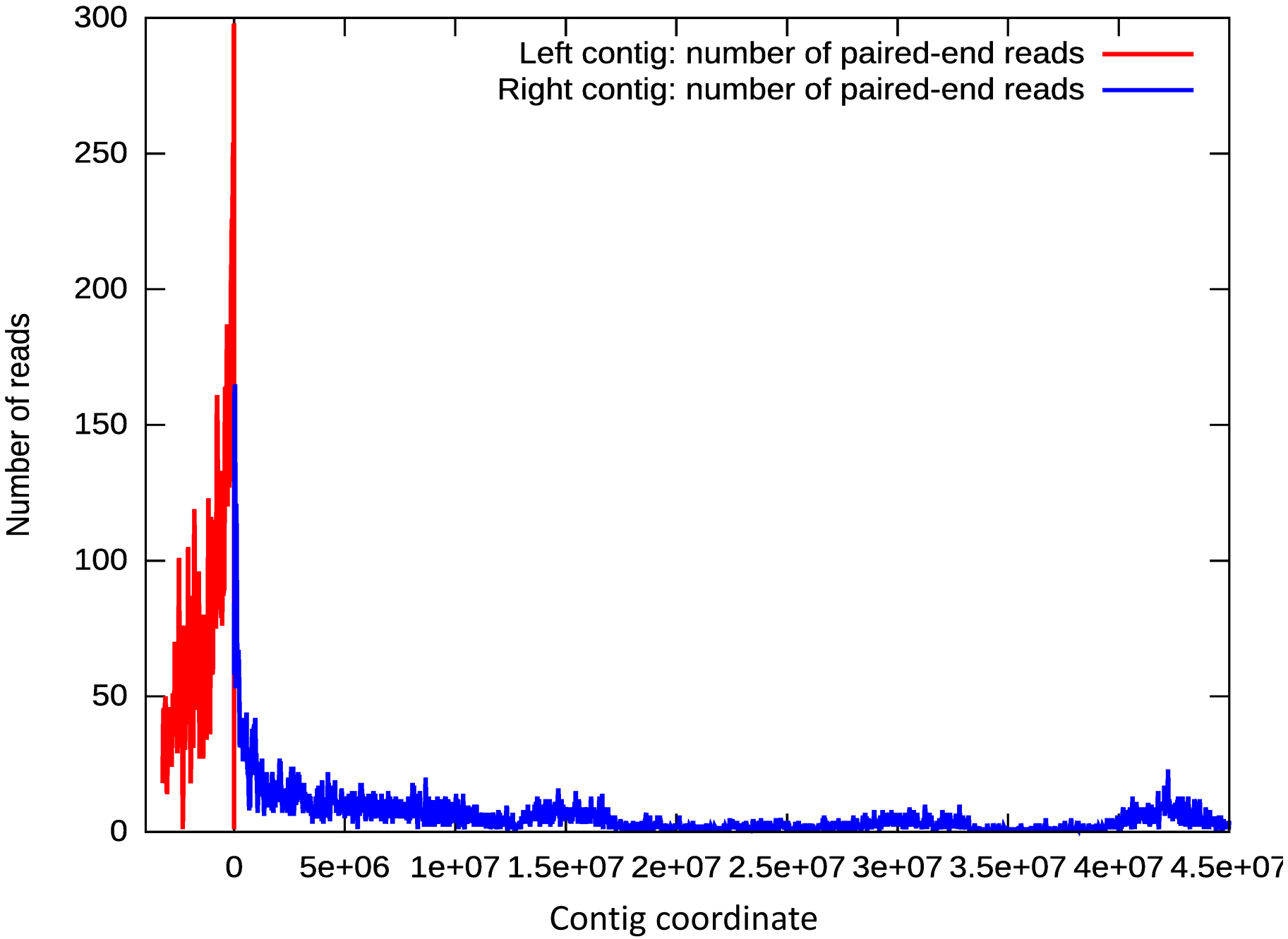




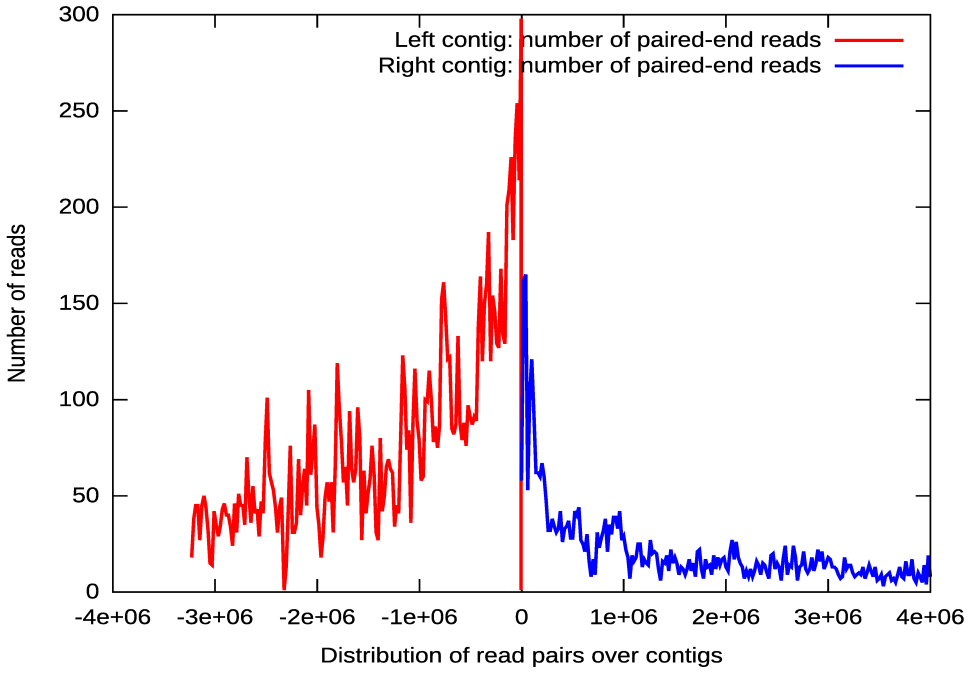
	In_Pairs	Different_Chr	No_hits
Human:	86.0%	13.9%	<0.3 %

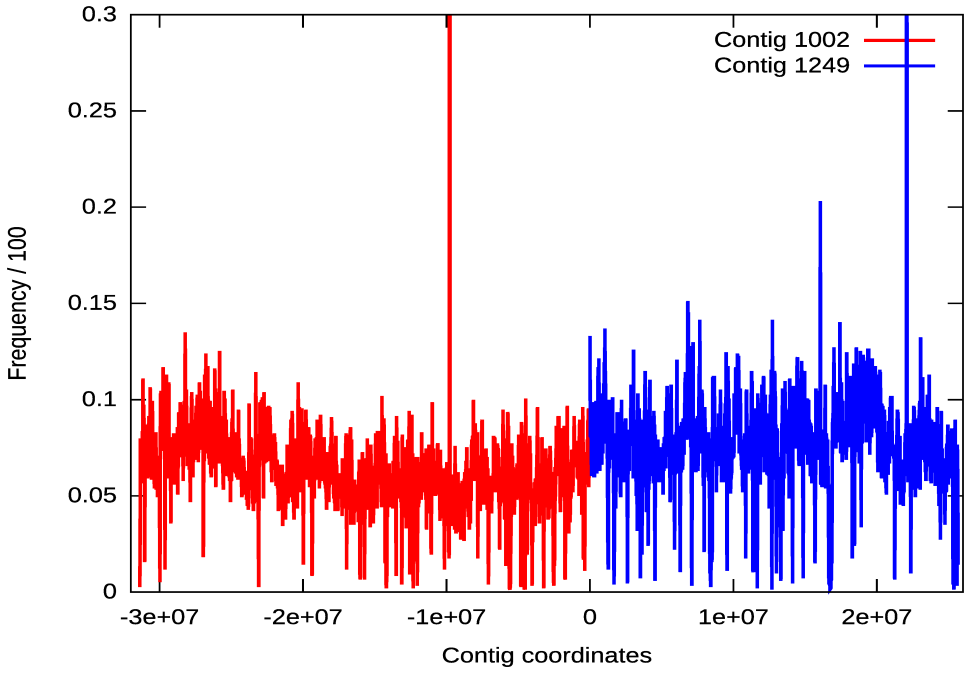


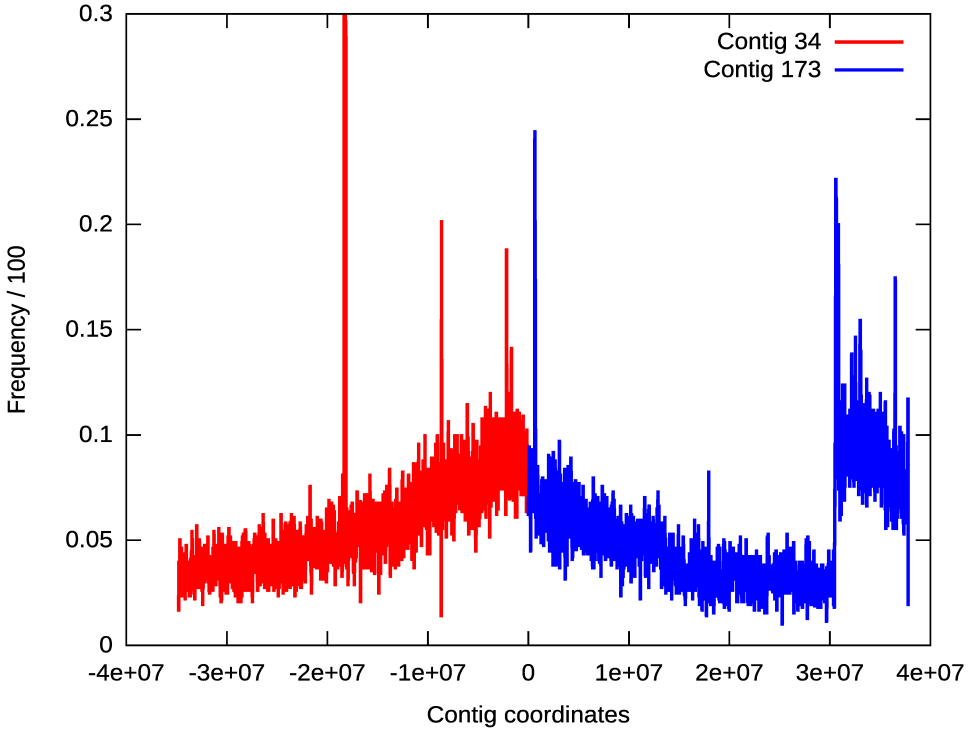
Pair-end Reads Distributions over Contig i and j

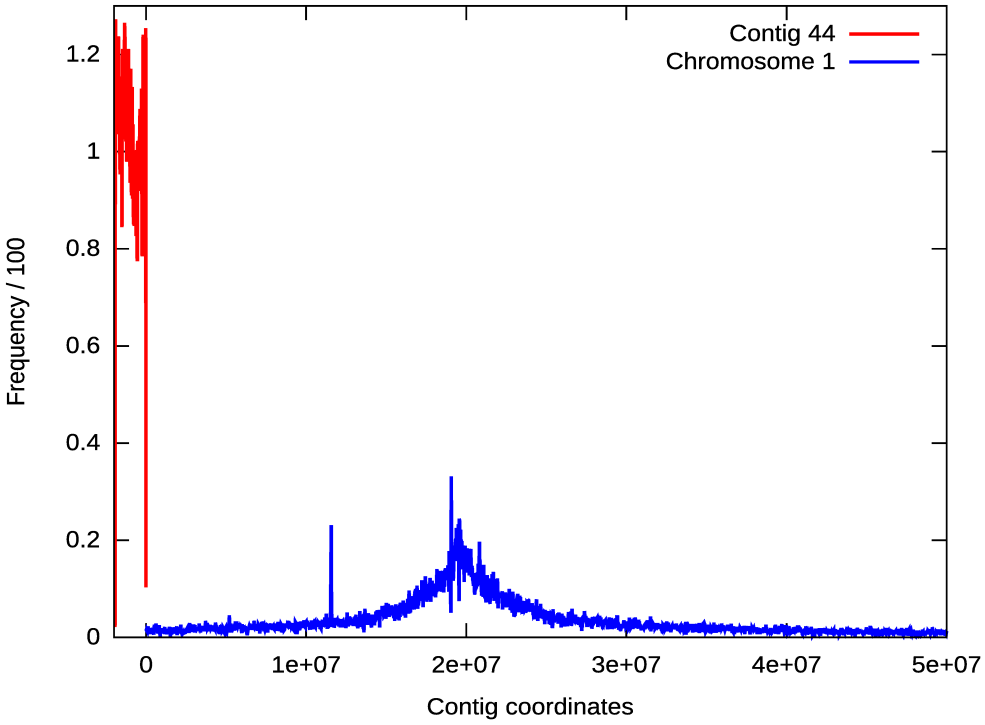


Distribution, Distribution and Distribution



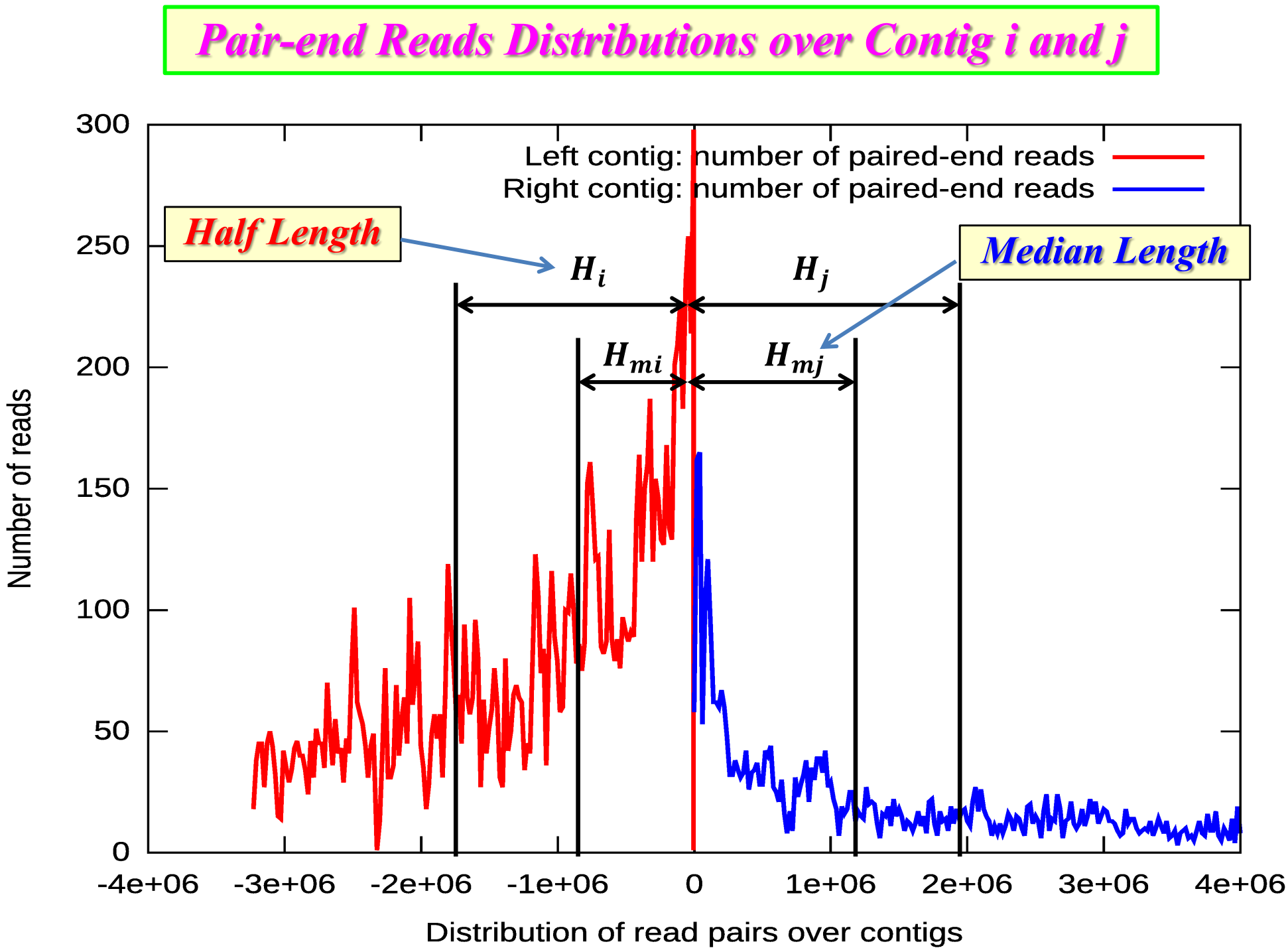






Contact Matrix: C(i,j) – number of paired-end reads between contig i and contig j

	1	2	3	4	5	6	… j …	N
1		4005	200	0	3000	0		
2	4005		870	0	0	0		
3	200	870		0	7400	0		
4	0	0	0		0	13560		
5	3000	0	7400	0			0	
6	0	0	0	13560	0			
i								
N							C(i,j)	


Pair-end Reads Distributions over Contig i and j
Left contig: number of paired-end reads
Right contig: number of paired-end reads
Half Length
Median Length
H_i
H_j
H_{mi}
H_{mj}
Number of reads
Distribution of read pairs over contigs
300
250
200
150
100
50
0
-4e+06
-3e+06
-2e+06
-1e+06
0
1e+06
2e+06
3e+06
4e+06

I_{CD} - Index of Contig Distance i and j

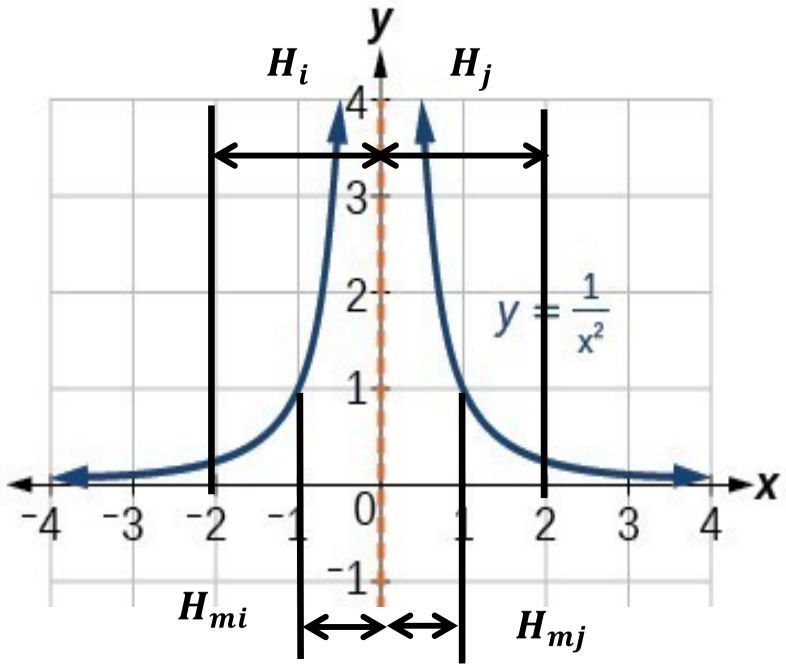
$$I_{CD} = \frac{GN_i}{N_T H_{mi}}\left(\frac{H_i}{H_{mi}} - 1\right) * \frac{GN_j}{N_T H_{mj}}\left(\frac{H_j}{H_{mj}} - 1\right)$$

Where

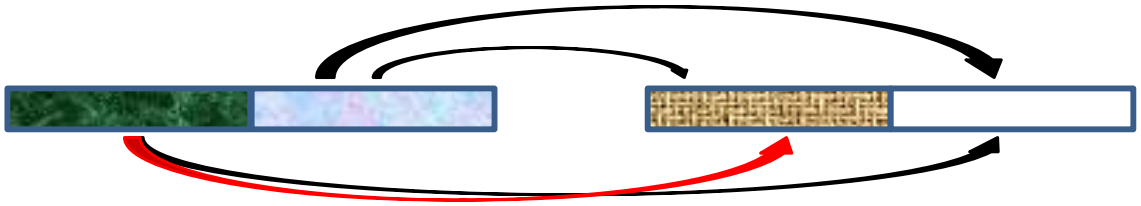
$G - Genome\ size;$
$N_T - Total\ number\ of\ read\ pairs;$
$N_i - Number\ of\ paired\ ends\ on\ Contig\ i;$
$N_j - Number\ of\ paired\ ends\ on\ Contig\ j;$
$H_i - Half\ length\ of\ Contig\ i;$
$H_j - Half\ length\ of\ Contig\ j;$
$H_{mi} - Median\ length\ of\ Contig\ i;$
$H_{mj} - Median\ length\ of Contig\ j;$

$\frac{H_i}{H_{mi}}\ and \frac{H_j}{H_{mj}} - length\ distribution\ ratio$

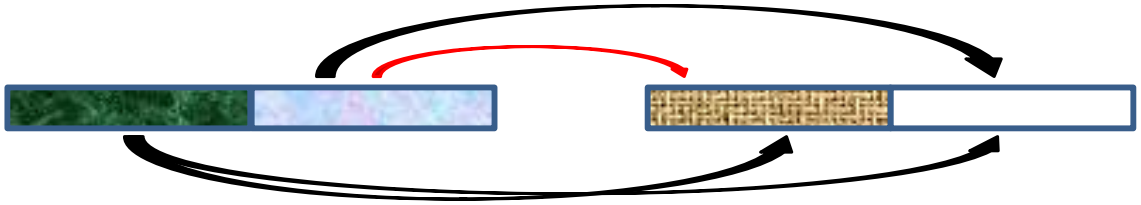
When
$H_i \to H_{mi}\ or\ H_j \to H_{mj}\ ,\quad I_{CD} -> 0$
$H_i \gg H_{mi}\ or H_j \gg H_{mj}\ ,\quad I_{CD} -> \infty$



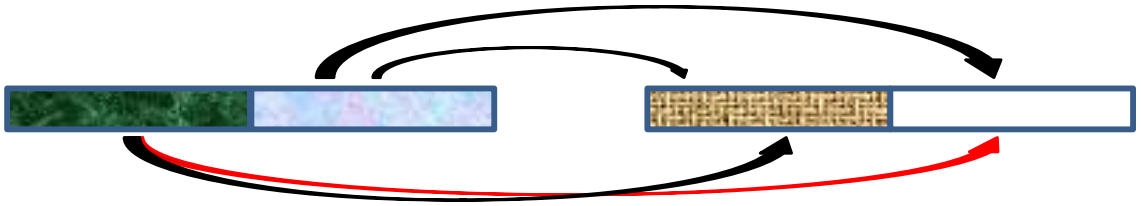
Contig Order and Orientation



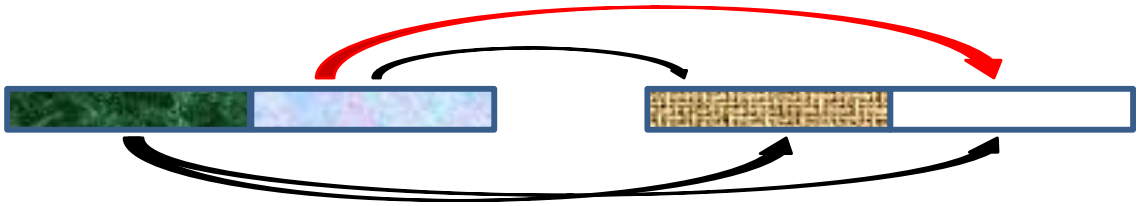
0: 2C-1F



0: 1F-2F



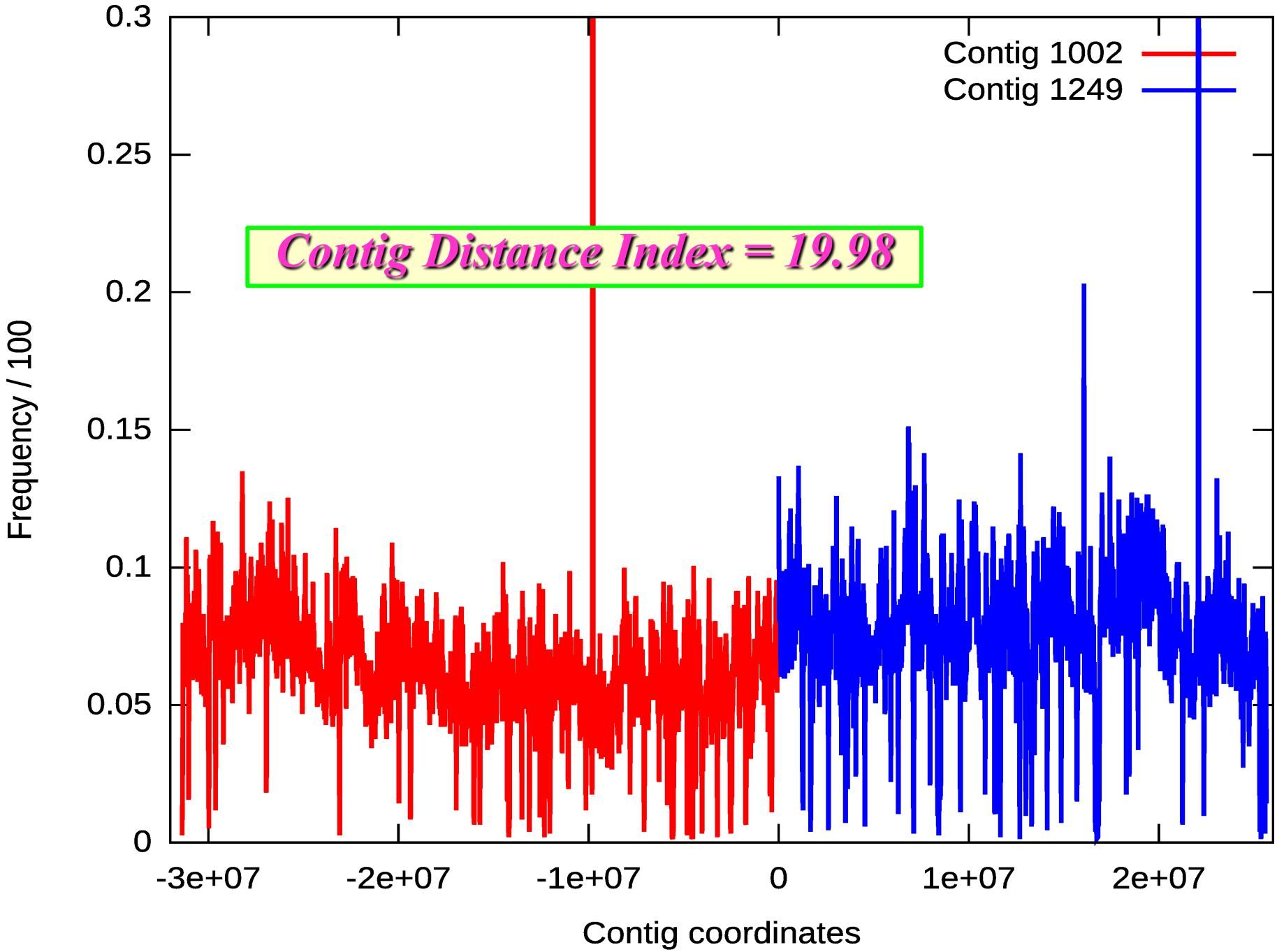
0: 2F-1F



0: 1F-2C

Contig Distance Index Identifies Best Partners

```
============================
scaffold: 1249 1468 25682287
============================
31401776 | 154098 1002 0 | 12582457 14135912 1.020559 1.110709 | 2.130245 9.381758 19.985445 38271 32359 37230 46238
28604069 | 136145 1125 1 | 11595245 13396149 1.107449 1.067623 | 9.836222 5.558078 54.670486 29919 42731 31747 31748
28710778 | 134191 0 0 | 11853917 13205683 1.083283 1.087061 | 7.514537 7.026868 52.803661 30791 29307 31363 42730
20597163 | 100413 273 0 | 12008208 9214524 1.069364 1.117647 | 4.683247 9.904203 46.383827 20758 24211 26286 29158
21243757 |  98636 1372 1 | 11670253 10329838 1.100331 1.028272 | 6.654194 2.266791 15.083669 21542 29461 23198 24435
20782805 |  95786 519 1 | 11730797 10367366 1.094652 1.002318 | 6.096165 0.184519 1.124857 20968 27030 22778 25010
16376805 |  87416 27 1 | 11907044 7651531 1.078449 1.070165 | 4.611092 6.467568 29.822554 19764 27234 20879 19539
17158708 |  83728 396 1 | 12407252 8040027 1.034971 1.067080 | 1.968792 5.652474 11.128545 19632 24905 20868 18323
15632158 |  83244 133 0 | 11776938 7529857 1.090364 1.038012 | 5.057898 3.495495 17.679855 19520 20246 18739 24739
15239653 |  82833 976 1 | 11624098 6546372 1.104700 1.163977 | 5.831429 15.391068 89.751915 19374 27119 18096 18244
14968830 |  80527 940 0 | 11255456 6899588 1.140882 1.084763 | 7.628163 7.874363 60.066929 15784 21009 19278 24456
14998092 |  80417 851 1 | 12001522 6885072 1.069960 1.089175 | 3.782846 8.256783 31.234140 19045 25040 18679 17653
14550196 |  78954 1027 1 | 11593829 6961076 1.107584 1.045111 | 5.711462 4.227145 24.143179 17369 23445 18134 20006
15622685 |  76561 952 1 | 11363457 7324850 1.130038 1.066417 | 6.694270 5.620649 37.626141 18061 22510 15782 20208
15940022 |  73676 150 3 | 12701833 7563450 1.010968 1.053753 | 0.543334 4.290428 2.331136 18271 20202 18943 16260
13589506 |  71922 964 0 | 11425098 6166075 1.123942 1.101958 | 5.993812 9.318269 55.851948 15172 16987 16857 22906
15322559 |  64734 728 2 | 12705148 7139090 1.010704 1.073145 | 0.465907 5.336343 2.486241 16442 13116 16266 18910
13683460 |  63785 605 1 | 10951173 6522470 1.172582 1.048948 | 7.401793 3.940148 29.164164 14390 19227 12462 17706
10973893 |  57336 1051 1 | 11297001 5291312 1.136606 1.036073 | 5.266470 3.335030 17.568090 13312 16579 11775 15670
1898579 |  50299 1396 0 | 7461457 747616 1.720997 1.269755 | 24.384659 123.412109 3009.362061 6293 13392 8690 21924
8702427 |  44822 1099 1 | 12476550 4238496 1.029222 1.026594 | 0.880702 2.365305 2.083128 11136 11909 10628 11149
7265313 |  40134 1210 1 | 10900207 3486876 1.178064 1.041808 | 4.805216 3.988209 19.164204 8564 12515 8117 10938
7747250 |  38339 1174 2 | 11938069 3473092 1.075647 1.115325 | 1.950086 9.855371 19.218821 10238 7205 10254 10642
7794968 |  37137 1014 1 | 11443708 3511170 1.122114 1.110024 | 3.049269 9.051884 27.601625 9020 11827 7551 8739
7064720 |  36484 1162 1 | 11785771 3496998 1.089546 1.010112 | 2.196718 0.901788 1.980973 8272 10164 8532 9516
6796151 |  36402 1138 0 | 11377440 3332582 1.128650 1.019652 | 3.148891 1.817753 5.723907 7812 10015 8135 10440
6452564 |  35652 1075 0 | 11811225 3073863 1.087198 1.049585 | 2.090334 4.731116 9.889614 7827 9256 8591 9978
2911094 |  34548 1150 1 | 9200948 821867 1.395633 1.771025 | 9.190497 158.013000 1452.218018 8131 15214 4208 6995
6629460 |  34096 1039 2 | 12399480 2920214 1.035619 1.135098 | 0.816611 11.998667 9.798244 8773 6623 8820 9880
6222009 |  32588 1371 0 | 10894139 3074581 1.178720 1.011847 | 3.916118 1.071461 4.195969 6690 9402 6853 9643
6166754 |  29618 1124 3 | 12129274 3049301 1.058690 1.011175 | 1.168814 0.926842 1.083306 7599 7389 8019 6611
5595688 |  26325 988 0 | 10771318 2527833 1.192161 1.106815 | 3.401397 8.677722 29.516376 4860 6691 6106 8668
5588080 |  24527 1333 3 | 11880407 2594407 1.080867 1.076947 | 1.333648 5.832218 7.778129 7019 6127 6175 5206
675500 |  23887 629 0 | 7752433 6161 1.656402 0.000000 | 10.542783 -610.652222 -6437.973633  487  759 9183 13458
4726344 |  23875 1111 1 | 10851558 2257096 1.183346 1.046997 | 2.943322 4.099617 12.066493 5154 7354 4841 6526
4727819 |  23583 260 2 | 11870336 2200129 1.081784 1.074441 | 1.296859 6.412223 8.315750 5780 5015 6889 5899
3699761 |  19858 125 0 | 12021889 1765179 1.068147 1.047984 | 0.909924 4.447535 4.046919 4606 4885 4718 5649
3944518 |  18395 1063 1 | 11187545 1816048 1.147807 1.086017 | 1.828179 6.927044 12.663879 4410 5792 3536 4657
3044807 |  16299 1222 1 | 10788702 1310912 1.190240 1.161331 | 2.084905 14.913423 31.093071 3851 5737 2913 3798
545741 |  16136 998 3 | 2275230 197877 5.643888 1.378988 | 50.385044 193.505432 9749.779297 8081 1660 5063 1332
3983511 |  16089 1261 2 | 10197384 1988885 1.259259 1.001443 | 2.804698 0.100646 0.282281 4641 3387 4734 3327
2828887 |  14756 1087 1 | 10953221 1382482 1.172362 1.023119 | 1.710152 2.082441 3.561291 3297 4254 2922 4283
4421250 |  14714 1321 2 | 10187857 2162578 1.260436 1.022218 | 2.576654 1.276847 3.289993 4084 3069 4495 3066
2404967 |  13993 1285 0 | 11059561 1140377 1.161090 1.054461 | 1.515664 5.471977 8.293676 2905 3742 3017 4329
3100475 |  13924 1000 1 | 10929356 1305746 1.174922 1.187242 | 1.637695 14.521044 23.781044 3444 4919 2433 3128
2926209 |  12479 1297 2 | 10036198 1441906 1.279483 1.014701 | 2.345087 1.082655 2.538920 3484 2645 3859 2491
2425694 |  11869 420 1 | 12293648 1174907 1.044535 1.032292 | 0.355417 2.728534 0.969766 2973 3099 2741 3056
281384 |  11360 592 1 | 7682396 13505 1.671502 0.000000 | 5.129203 -697.166748 -3575.909912 4546 6330  227  257
2516334 |  11172 2 2 | 12404780 1255190 1.035177 1.002372 | 0.264249 0.181843 0.048052 2916 2661 2873 2722
2104551 |  10696 1444 1 | 12024683 1041658 1.067899 1.010192 | 0.488322 0.894532 0.436820 2664 2773 2361 2898
mapp: 1 1249 1396 50299 || 3009.362061 0 1 1.720997 1.269755
mapp: 2 1249 998 16136 || 9749.779297 3 1 5.643888 1.378988
```


Contig 1002
Contig 1249
Contig Distance Index = 19.98
Frequency / 100
0.3
0.25
0.2
0.15
0.1
0.05
0
-3e+07
-2e+07
-1e+07
0
1e+07
2e+07
Contig coordinates

```
====================
matrix: 1124 6166754
====================
14968830 | 116654 940 2 | 2503825 6126773 1.231467 1.221592 | 75.611732 29.821056 2254.821777 27701 18874 42467 27612
31401776 |  41411 1002 1 | 2996262 14675304 1.029075 1.069885 | 3.371551 1.591488 5.365782 10348 11726 9983 9354
28710778 |  32631 0 1 | 3070339 13718818 1.004246 1.046401 | 0.388025 0.910695 0.353372 8240 8843 8009 7539
28604069 |  31086 1125 2 | 3000242 13022402 1.027709 1.098264 | 2.412095 1.844118 4.448186 7665 6439 8356 8626
25682287 |  29618 1249 3 | 3049301 12129274 1.011175 1.058690 | 0.926842 1.168814 1.083306 7599 8019 7389 6611
20597163 |  24474 273 2 | 3052322 8287611 1.010174 1.242648 | 0.697284 4.978879 3.471691 5190 4674 7179 7431
21243757 |  22569 1372 2 | 2958711 9724937 1.042135 1.092231 | 2.662925 1.692060 4.505829 5612 4738 6103 6116
20782805 |  21976 519 2 | 2918322 9252896 1.056558 1.123043 | 3.480525 2.246776 7.819963 5336 4565 6232 5843
15632158 |  21813 133 3 | 3071218 7212020 1.003959 1.083757 | 0.241829 2.018258 0.488074 5736 5948 5227 4902
17158708 |  21047 396 2 | 3042006 8037360 1.013600 1.067434 | 0.801542 1.428382 1.144907 5202 4613 5483 5749
14998092 |  20995 851 2 | 3081896 7469535 1.000481 1.003951 | 0.028251 0.095505 0.002698 5383 5066 5122 5424
1369331 |  20736 518 1 | 1867207 507264 1.651331 1.349721 | 37.820553 91.452698 3458.791504 4141 8585 2568 5442
15940022 |  20373 150 0 | 3052112 7399459 1.010244 1.077107 | 0.584408 1.701842 0.994570 4956 4668 5120 5629
16376805 |  20210 27 0 | 2968328 7352280 1.038759 1.113723 | 2.193505 2.423497 5.315952 4777 4403 5044 5986
15239653 |  19730 976 3 | 3002917 7532553 1.026794 1.011586 | 1.480352 0.259028 0.383452 5203 4767 4943 4817
14550196 |  18938 1027 2 | 3079603 7015528 1.001225 1.036999 | 0.064989 0.831605 0.054045 4591 4519 4898 4930
15322559 |  18008 728 1 | 2879966 7211981 1.070630 1.062299 | 3.561667 1.264362 4.503237 4217 5266 4270 4255
15622685 |  16929 952 3 | 2887423 6810778 1.067865 1.146909 | 3.217185 2.749043 8.844179 4985 4579 3973 3392
13589506 |  16907 964 2 | 3028028 6564605 1.018279 1.035059 | 0.865399 0.753217 0.651833 4090 4040 4497 4280
13683460 |  12780 605 3 | 2897790 6257470 1.064044 1.093370 | 2.291990 1.505913 3.451537 3606 3330 3153 2691
10973893 |  12401 1051 3 | 2878187 5126732 1.071291 1.070262 | 2.475684 1.371115 3.394449 3525 3150 3072 2654
8702427 |  11869 1099 3 | 3067897 4105694 1.005046 1.059800 | 0.167704 1.408414 0.236196 3065 3210 2909 2685
7747250 |  11830 1174 1 | 2820309 3756928 1.093276 1.031062 | 3.089989 0.819070 2.530918 2557 3499 2858 2916
6629460 |  10143 1039 1 | 2880132 3261106 1.070568 1.016443 | 2.004357 0.434450 0.870793 2221 2930 2523 2469
7064720 |  9220 1162 2 | 3020717 3421696 1.020743 1.032342 | 0.535564 0.728884 0.390364 2324 2121 2384 2391
6452564 |  9013 1075 0 | 3058056 2949043 1.008280 1.094010 | 0.208979 2.267610 0.473883 2090 1985 2391 2547
7265313 |  8971 1210 3 | 2960136 3602719 1.041634 1.008310 | 1.045890 0.177184 0.185315 2317 2218 2328 2108
6796151 |  8541 1138 2 | 3042740 3259196 1.013355 1.042611 | 0.319423 0.924762 0.295390 2070 1978 2259 2234
7794968 |  8247 1014 3 | 3017528 3496476 1.021822 1.114689 | 0.503960 2.095376 1.055985 2381 2242 1840 1784
4727819 |  7753 260 0 | 2958219 2238838 1.042309 1.055864 | 0.918542 1.581978 1.453114 1832 1806 1904 2211
5588080 |  7716 1333 1 | 2927870 2552813 1.053113 1.094495 | 1.147601 2.253171 2.585743 1989 2224 1692 1811
6222009 |  6626 1371 3 | 2947755 3061110 1.046009 1.016299 | 0.853673 0.299743 0.255883 1749 1598 1713 1566
3983511 |  5281 1261 1 | 2620471 1921466 1.176650 1.036581 | 2.612346 0.837457 2.187728 1225 1476 1072 1508
5595688 |  5278 988 2 | 2908449 2462869 1.060145 1.136010 | 0.888932 2.215362 1.969306 1227 1056 1569 1426
675500 |  5275 629 0 | 2208641 6059 1.396052 0.000000 | 5.850263 -134.851196 -788.914917  105  119 1832 3219
4421250 |  5222 1321 0 | 2695543 2116552 1.143880 1.044446 | 2.103961 0.906537 1.907319 1155 1327 1128 1612
3699761 |  4983 125 3 | 3076064 1838542 1.002377 1.006167 | 0.033174 0.143429 0.004758 1260 1247 1238 1238
4726344 |  4736 1111 3 | 2777960 2312817 1.109943 1.021772 | 1.458076 0.376743 0.549320 1295 1127 1305 1009
2926209 |  4288 1297 0 | 2493551 1414241 1.236541 1.034551 | 2.840282 0.874307 2.483277  866 1199  881 1342
2470652 |  3782 1457 1 | 2461134 1185057 1.252828 1.042419 | 2.677610 1.121318 3.002453  780 1206  777 1019
3944518 |  3761 1063 3 | 2880049 1775042 1.070599 1.111106 | 0.743536 1.829376 1.360207 1125 1008  878  750
2516334 |  3447 2 0 | 3066901 1226064 1.005372 1.026184 | 0.051855 0.619390 0.032119  872  807  844  924
3044807 |  3260 1222 3 | 2805927 1301318 1.098880 1.169893 | 0.902666 3.141171 2.835427 1040  898  740  582
2438354 |  3196 272 0 | 2615448 1086565 1.178910 1.122047 | 1.601185 2.762451 4.423197  595  812  801  988
2404967 |  3172 1285 2 | 3065329 1154190 1.005888 1.041842 | 0.052298 0.952992 0.049839  747  790  847  788
2425694 |  3160 420 1 | 3068889 1096689 1.004721 1.105917 | 0.041775 2.382728 0.099538  840  871  734  715
2828887 |  3092 1087 3 | 3073616 1338280 1.003176 1.056911 | 0.027497 1.074184 0.029537  820  810  734  728
1904542 |  2875 433 1 | 2805822 894892 1.098921 1.064118 | 0.796394 1.671428 1.331114  748  819  591  717
2202572 |  2824 408 1 | 2899636 1027032 1.063367 1.072300 | 0.501104 1.600767 0.802151  707  795  632  690
2351542 |  2767 1359 0 | 2503130 1032689 1.231809 1.138553 | 1.796138 2.815333 5.056726  511  669  601  986
mapp: 1 1124 518 20736 || 3458.791504 1 0 1.651331 1.349721
mapp: 2 1124 940 116654 || 2254.821777 2 0 1.231467 1.221592
```

Partners and Partnership

(1) A partnership is unique;

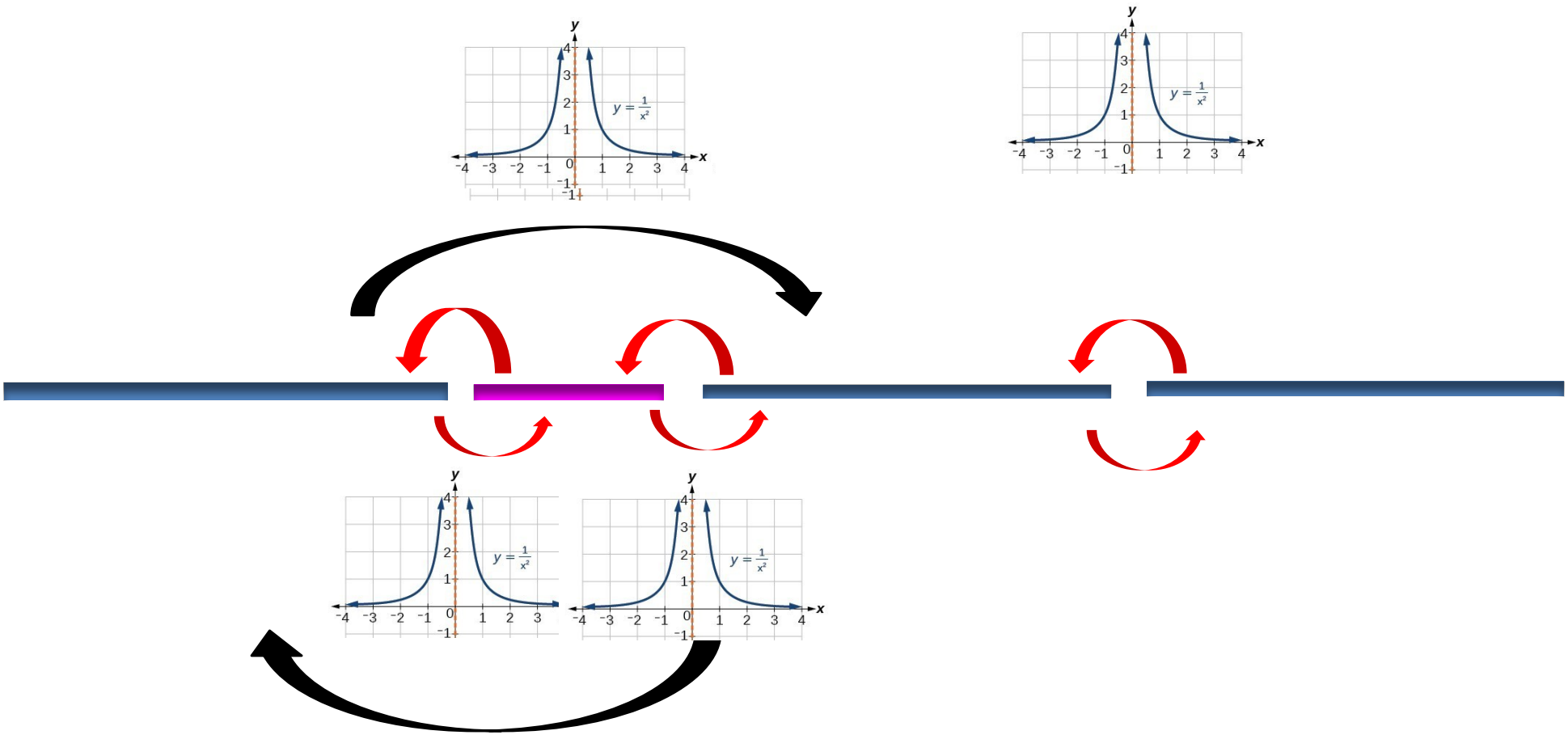
(2) Each partner is independent;

(3) Establishment of a partnership is determined by the willingness of both sides:

(a) Unique mutual best partner

(b) Select partner with a criteria when there are choices

Re-adjust the Best Partner



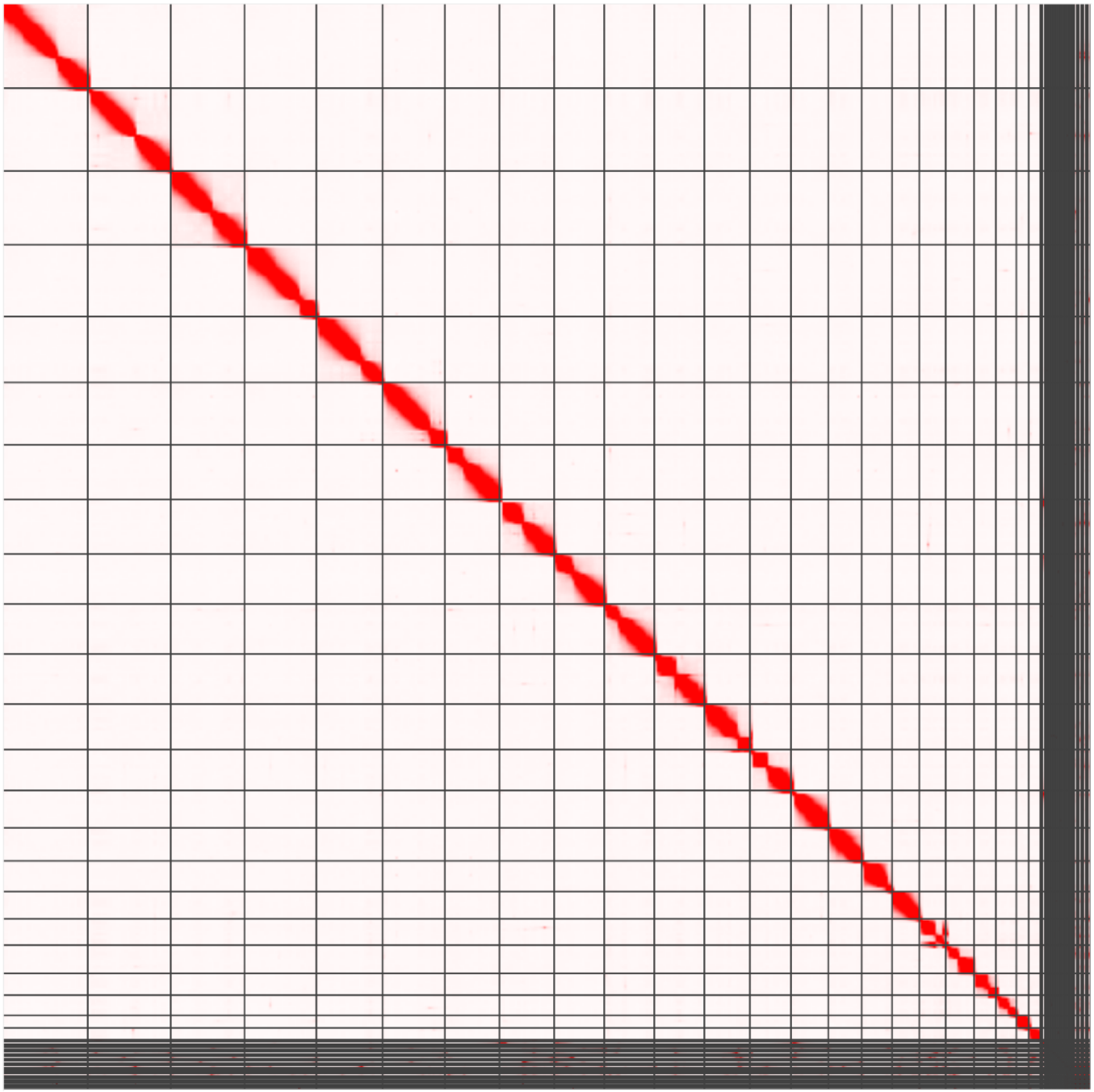
1 2 3 4

Partner Matrix: P(i,j) – number of paired-end reads between contig i and contig j

	1	2	3	4	5	6	… j …	N
1		$P_{1,2}$	0	0	0	0		
2	$P_{2,1}$		$P_{2,3}$	0	0	0		
3	0	$P_{3,2}$		0	$P_{3,5}$	0		
4	0	0	0		0	$P_{4,6}$		
5	0	0	$P_{5,3}$	0		0		
6	0	0	0	$P_{6,4}$	0			
i								
N								

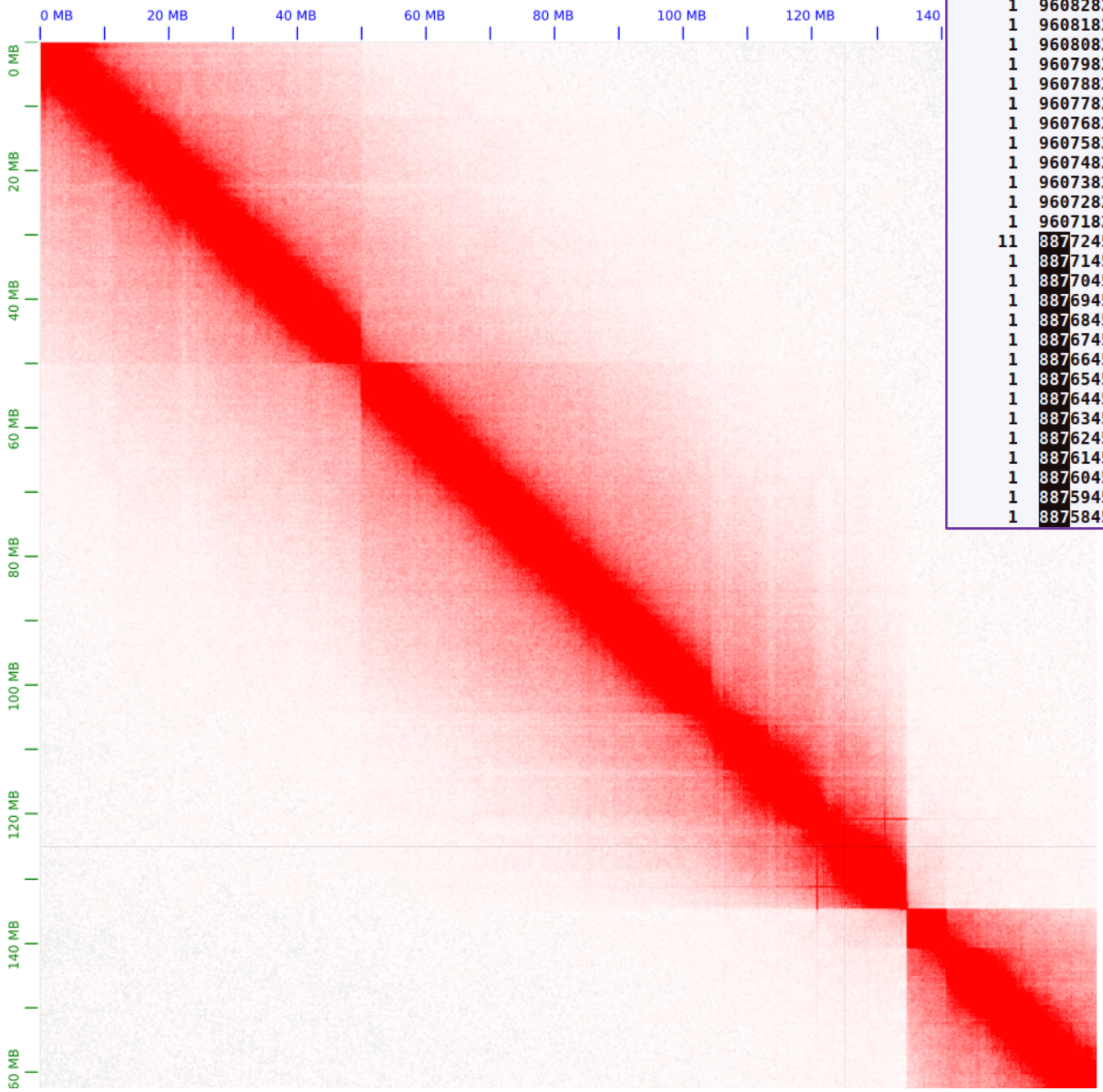
P(i,j)

Human HG002 Assembly



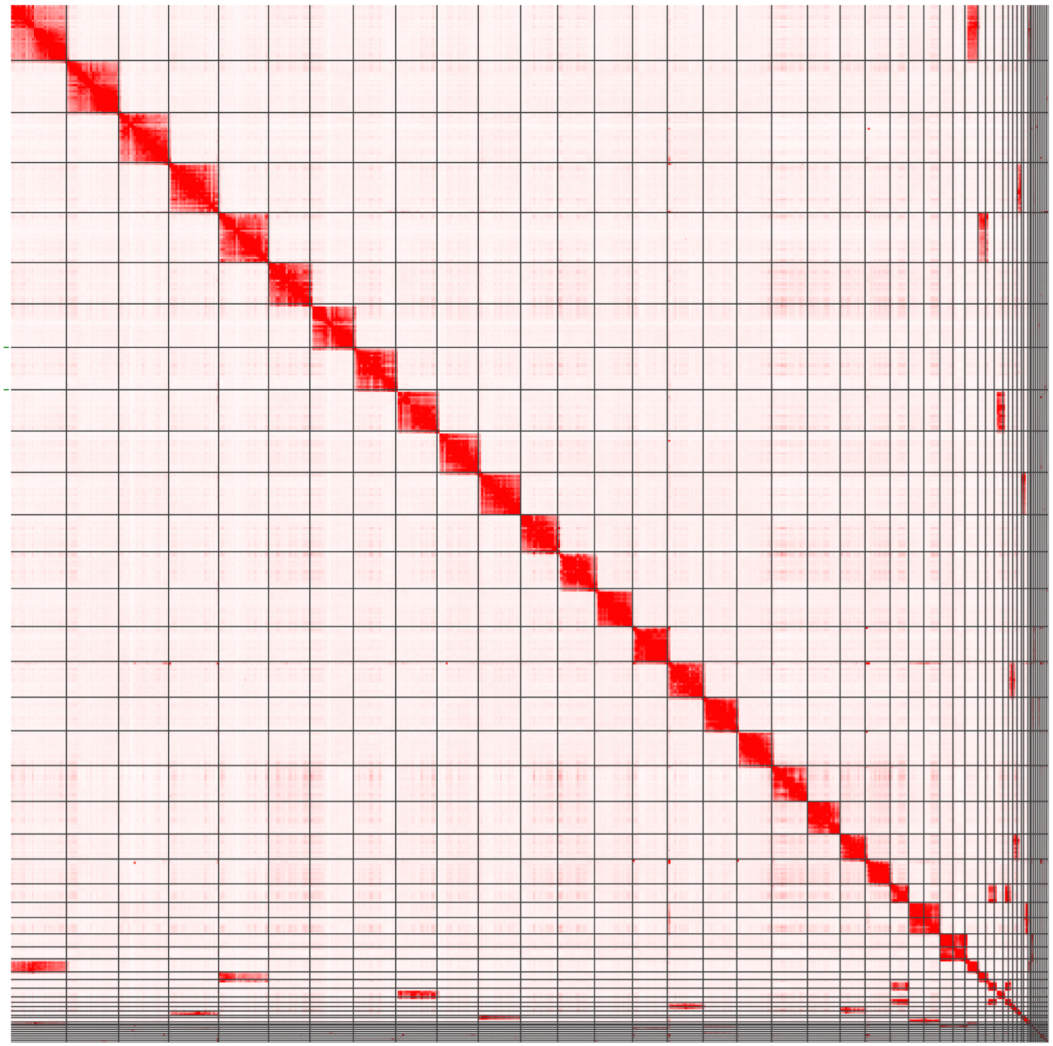
	WTdbg	*ScaffHiC*
Bases (Mb)	2732	2732
Number scaffolds	2213	1762
Scaffold N50 (Mb)	11.2	134
Scaffold N90 (Mb)	1.4	58
Max scaffold (Mb)	71.7	218

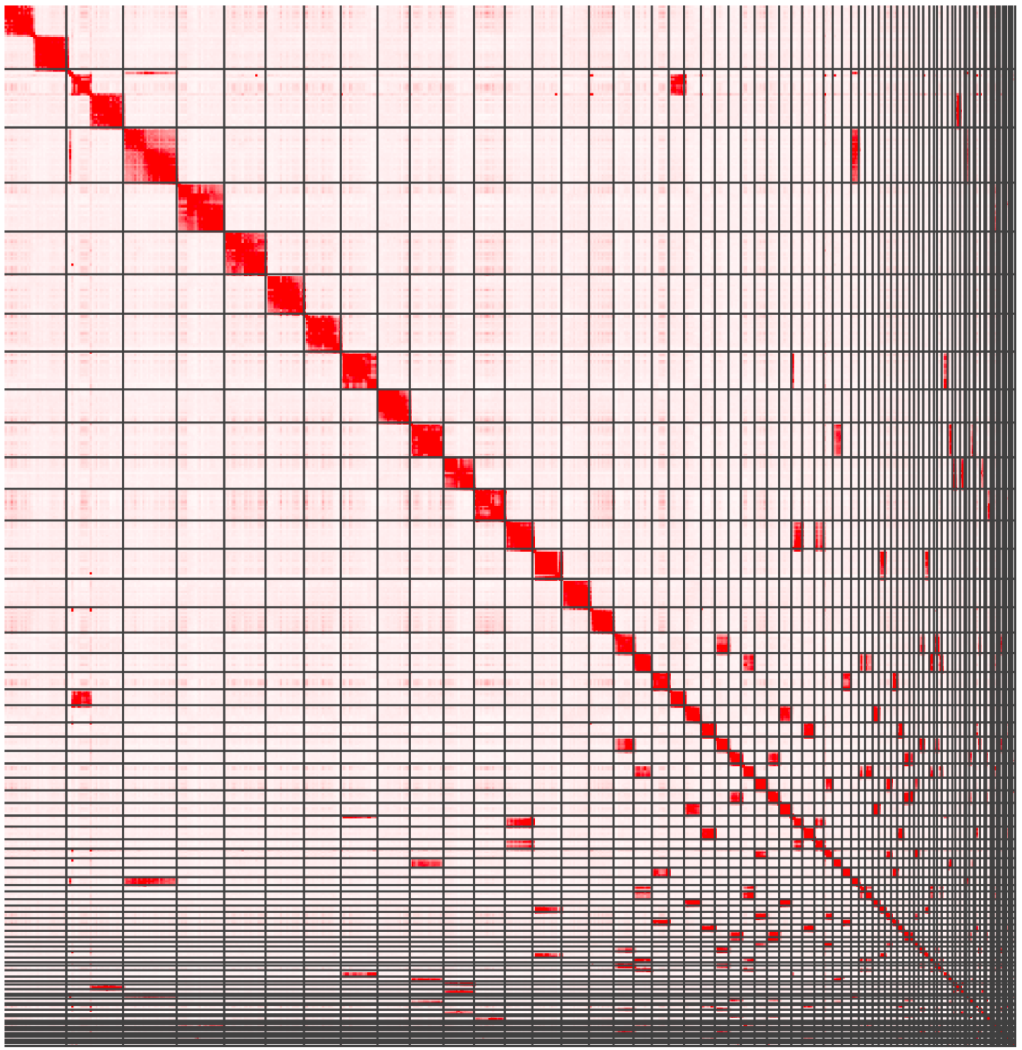
Chromosome 2



1	96086954	96087969	C	999	98.33	1000 242193529
1	96085954	96086953	C	1000	100.00	1000 242193529
1	96085250	96085953	C	704	100.00	1000 242193529
1	96083823	96084822	C	1000	100.00	1000 242193529
1	96082823	96083822	C	1000	100.00	1000 242193529
1	96081823	96082822	C	1000	100.00	1000 242193529
1	96080823	96081822	C	1000	100.00	1000 242193529
1	96079823	96080822	C	1000	100.00	1000 242193529
1	96078823	96079822	C	1000	100.00	1000 242193529
1	96077823	96078822	C	999	99.90	1000 242193529
1	96076823	96077822	C	1000	100.00	1000 242193529
1	96075823	96076822	C	1000	100.00	1000 242193529
1	96074823	96075822	C	1000	100.00	1000 242193529
1	96073823	96074822	C	1000	100.00	1000 242193529
1	96072823	96073822	C	999	99.90	1000 242193529
1	96071823	96072822	C	1000	100.00	1000 242193529
11	88772456	88773445	C	789	79.70	1000 242193529
1	88771456	88772455	C	1000	100.00	1000 242193529
1	88770455	88771455	C	1000	99.90	1000 242193529
1	88769455	88770454	C	1000	100.00	1000 242193529
1	88768455	88769454	C	999	99.90	1000 242193529
1	88767455	88768454	C	1000	100.00	1000 242193529
1	88766455	88767454	C	1000	100.00	1000 242193529
1	88765455	88766454	C	1000	100.00	1000 242193529
1	88764455	88765454	C	1000	100.00	1000 242193529
1	88763455	88764454	C	1000	100.00	1000 242193529
1	88762455	88763454	C	999	99.90	1000 242193529
1	88761455	88762454	C	1000	100.00	1000 242193529
1	88760455	88761454	C	1000	100.00	1000 242193529
1	88759455	88760454	C	1000	100.00	1000 242193529
1	88758454	88759454	C	1000	99.90	1000 242193529

fSclFor1: ScaffHiC vs Salsa2

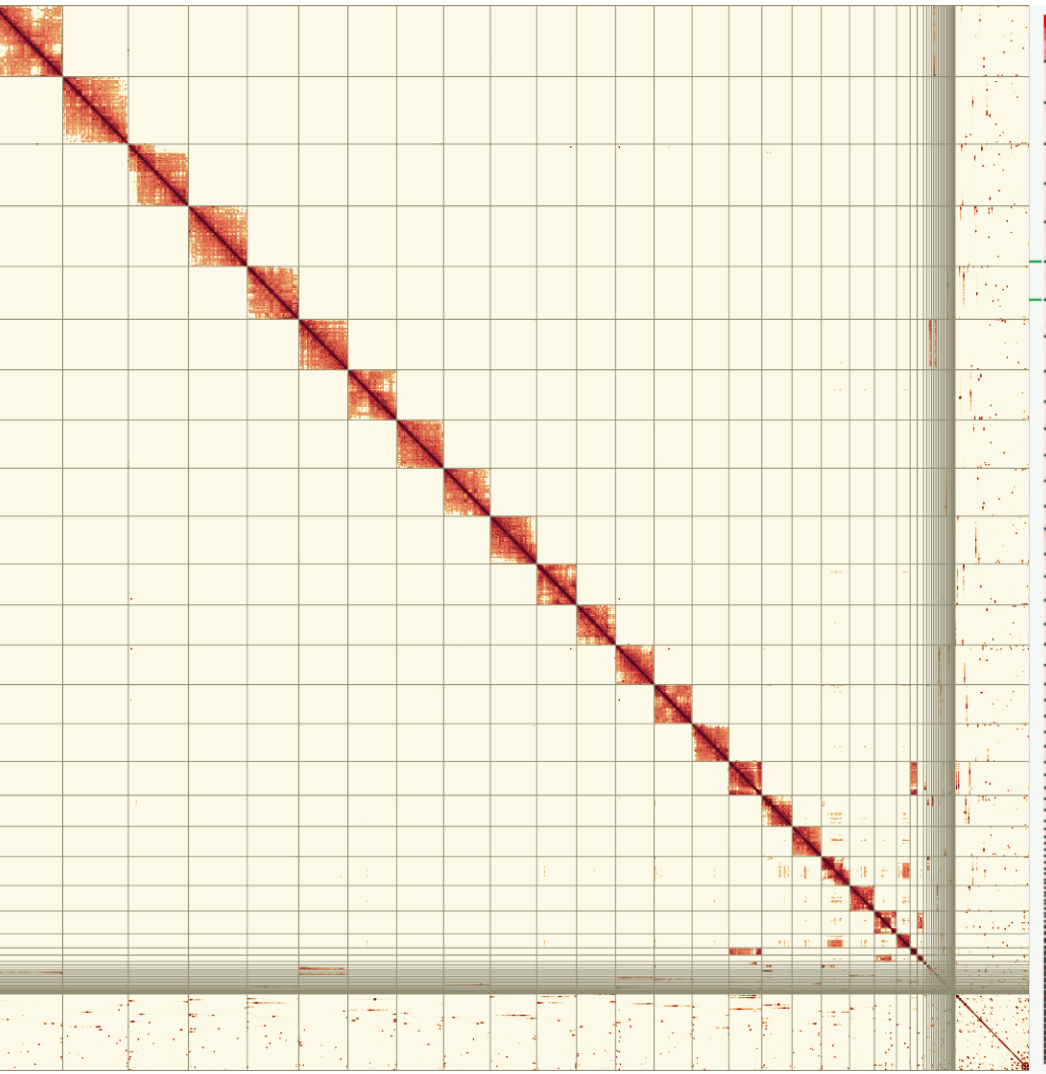


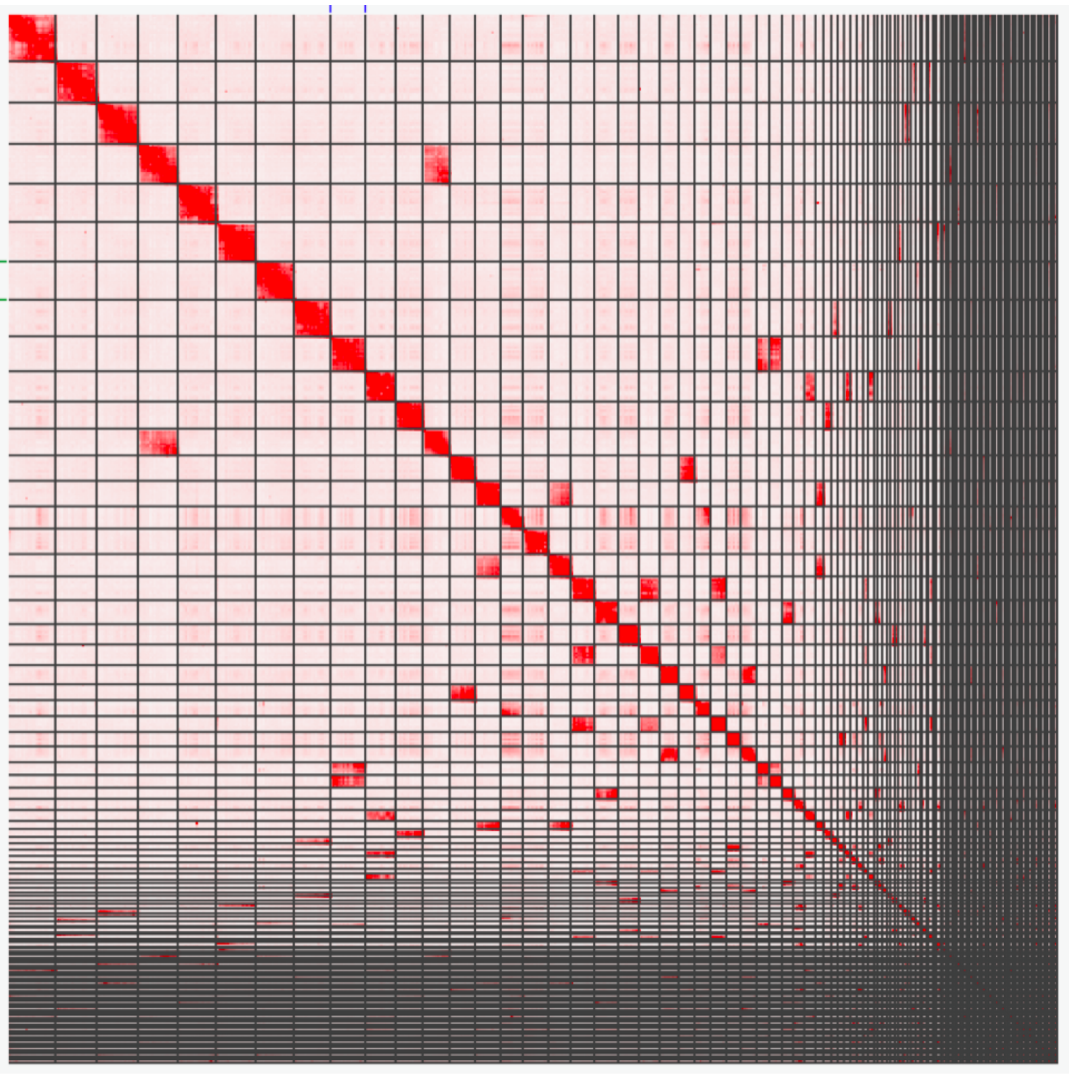


```
========
ScaffHiC
========
sum = 785500277, n = 123, ave = 6386181
      largest = 43024501
N50 = 29668867, n = 12
N60 = 27165132, n = 15
N70 = 25967964, n = 18
N80 = 19888780, n = 21
N90 = 10321926, n = 26
N100 = 2021, n = 123
```

```
======
SALSA2
======
sum = 785521263, n = 123, ave = 6386351
      largest = 50076388
N50 = 22561231, n = 13
N60 = 18658253, n = 16
N70 = 11901779, n = 22
N80 = 8474493, n = 31
N90 = 4017783, n = 45
N100 = 23590, n = 123
```

fThaAma1: ScaffHiC vs Salsa2





Summary

- ***The Contig Distance Index based on empirical mathematical modelling works well and there are hardly any trans-chromosome errors***
- ***It is possible to generate chromosome level and error free human genome assemblies***
- ***More work is need to sort out intra-chromosome rearrangements***
- ***PacBio, 10X, Bionano and HiC:***

Do we need all 4 data types?

Acknowledgements:

- *Shane MaCarthy*
- *Edward Harry*
- *Ying Yan*
- *Jo Wood*
- *Will Chow*
- *Kerstin Howe*
- *David Jackson*
- *Mike Quail*
- *Richard Durbin*

Phase Genomics