SeqOccIn

INRA
SCIENCE & IMPACT

Axis 1
Genome assembly

Sequençage Occitanie Innovation
2019-2021

Genotoul GeT

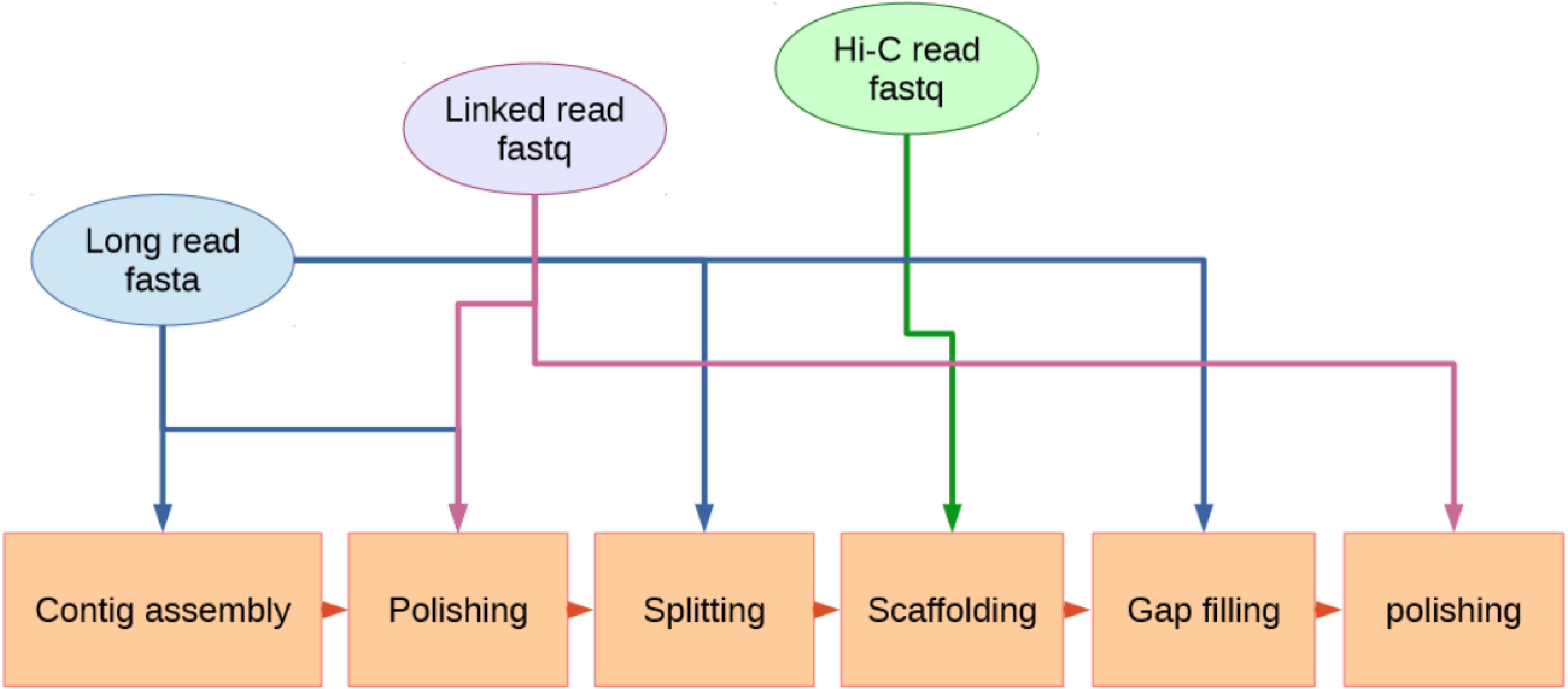
http://get.genotoul.fr
@get_genotoul

http://bioinfo.genotoul.fr

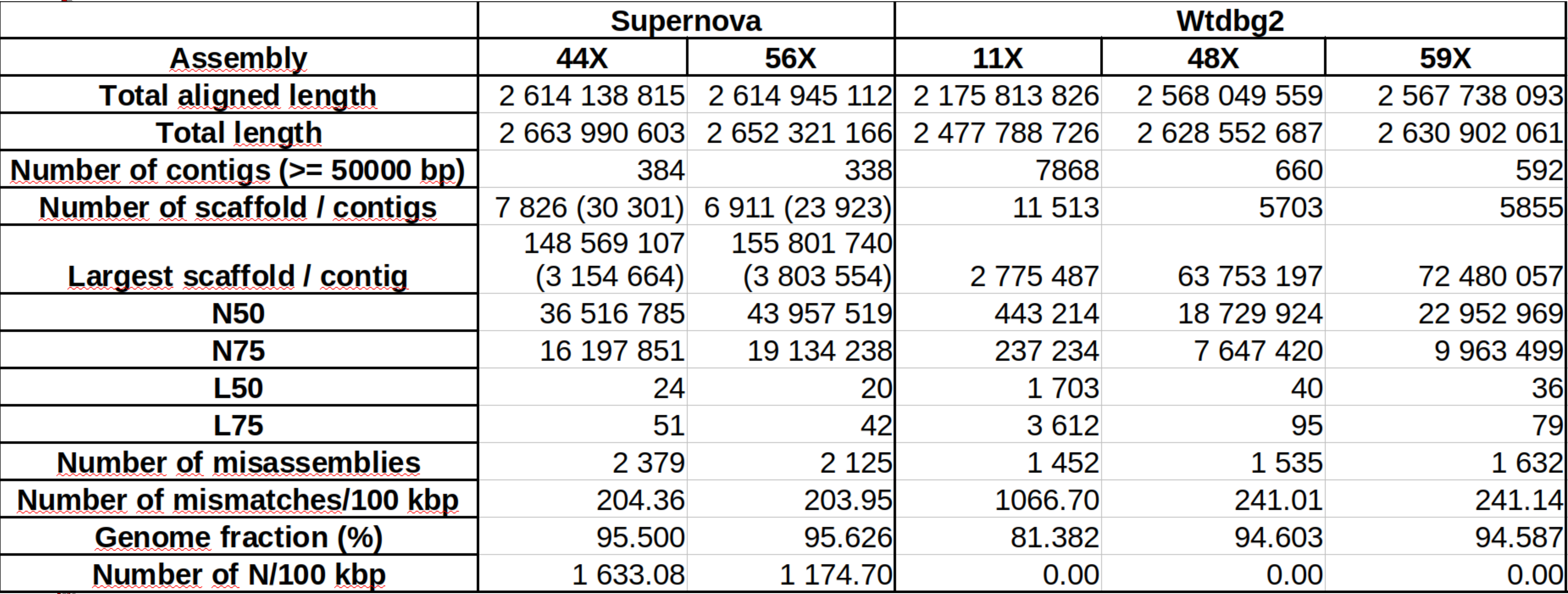
Genotoul Bioinfo

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

The pipeline layout

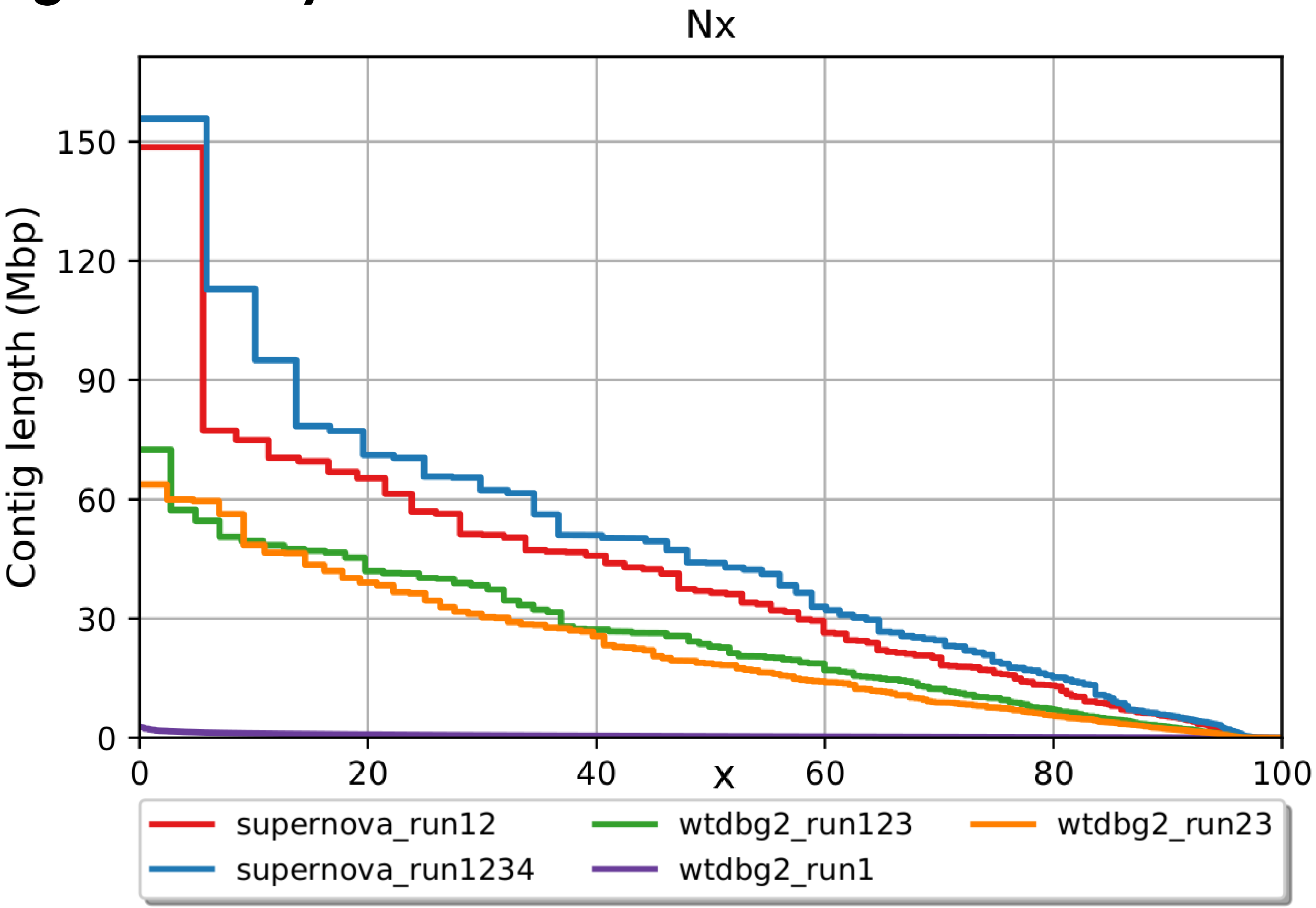


Contig assembly

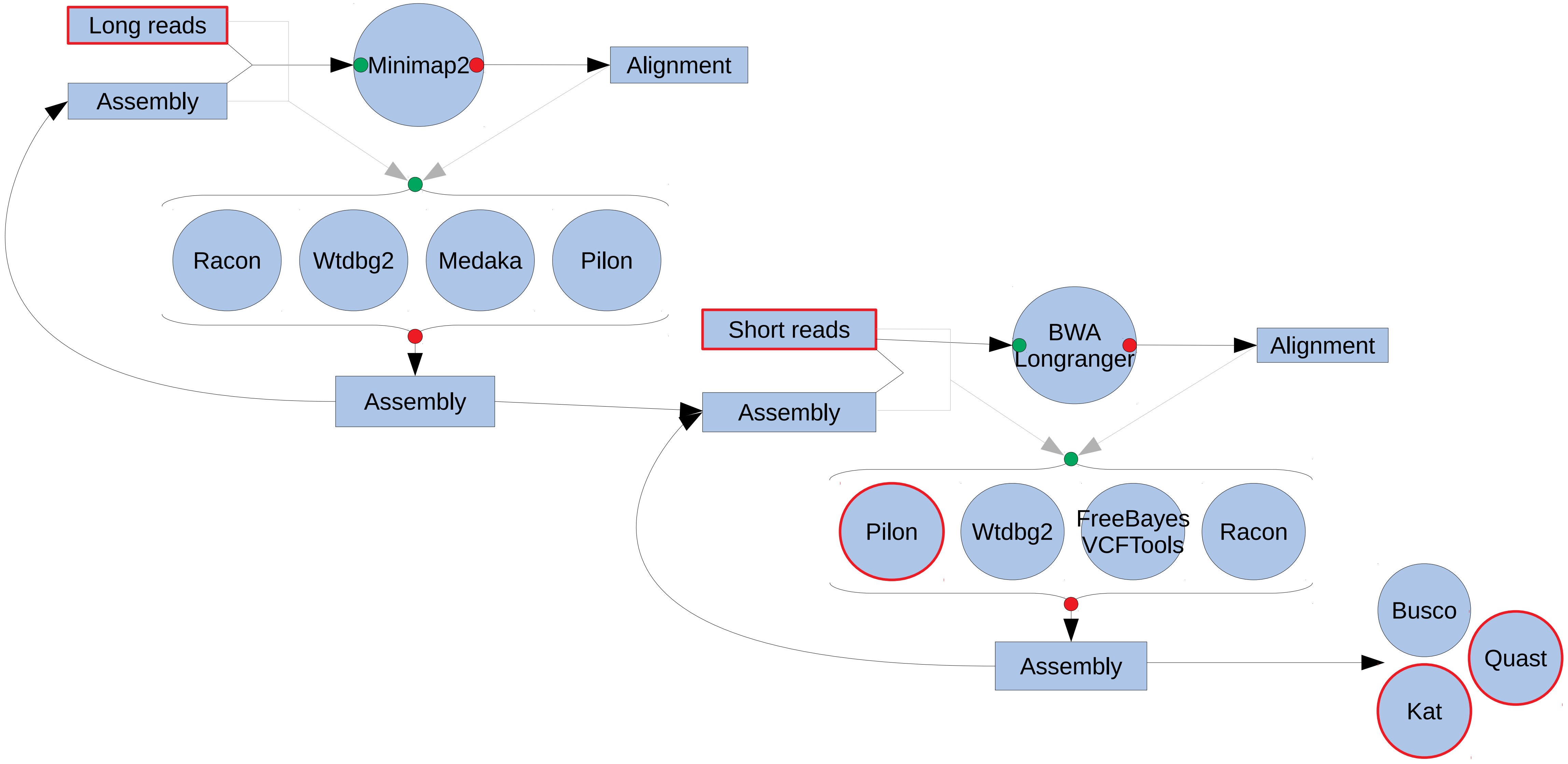


	Supernova		Wtdbg2		
Assembly	44X	56X	11X	48X	59X
Total aligned length	2 614 138 815	2 614 945 112	2 175 813 826	2 568 049 559	2 567 738 093
Total length	2 663 990 603	2 652 321 166	2 477 788 726	2 628 552 687	2 630 902 061
Number of contigs (>= 50000 bp)	384	338	7868	660	592
Number of scaffold / contigs	7 826 (30 301)	6 911 (23 923)	11 513	5703	5855
Largest scaffold / contig	148 569 107 (3 154 664)	155 801 740 (3 803 554)	2 775 487	63 753 197	72 480 057
N50	36 516 785	43 957 519	443 214	18 729 924	22 952 969
N75	16 197 851	19 134 238	237 234	7 647 420	9 963 499
L50	24	20	1 703	40	36
L75	51	42	3 612	95	79
Number of misassemblies	2 379	2 125	1 452	1 535	1 632
Number of mismatches/100 kbp	204.36	203.95	1066.70	241.01	241.14
Genome fraction (%)	95.500	95.626	81.382	94.603	94.587
Number of N/100 kbp	1 633.08	1 174.70	0.00	0.00	0.00

Contig assembly



Polishing pipeline



Polishing pipeline

Finished on Mother1_37165 subsampled

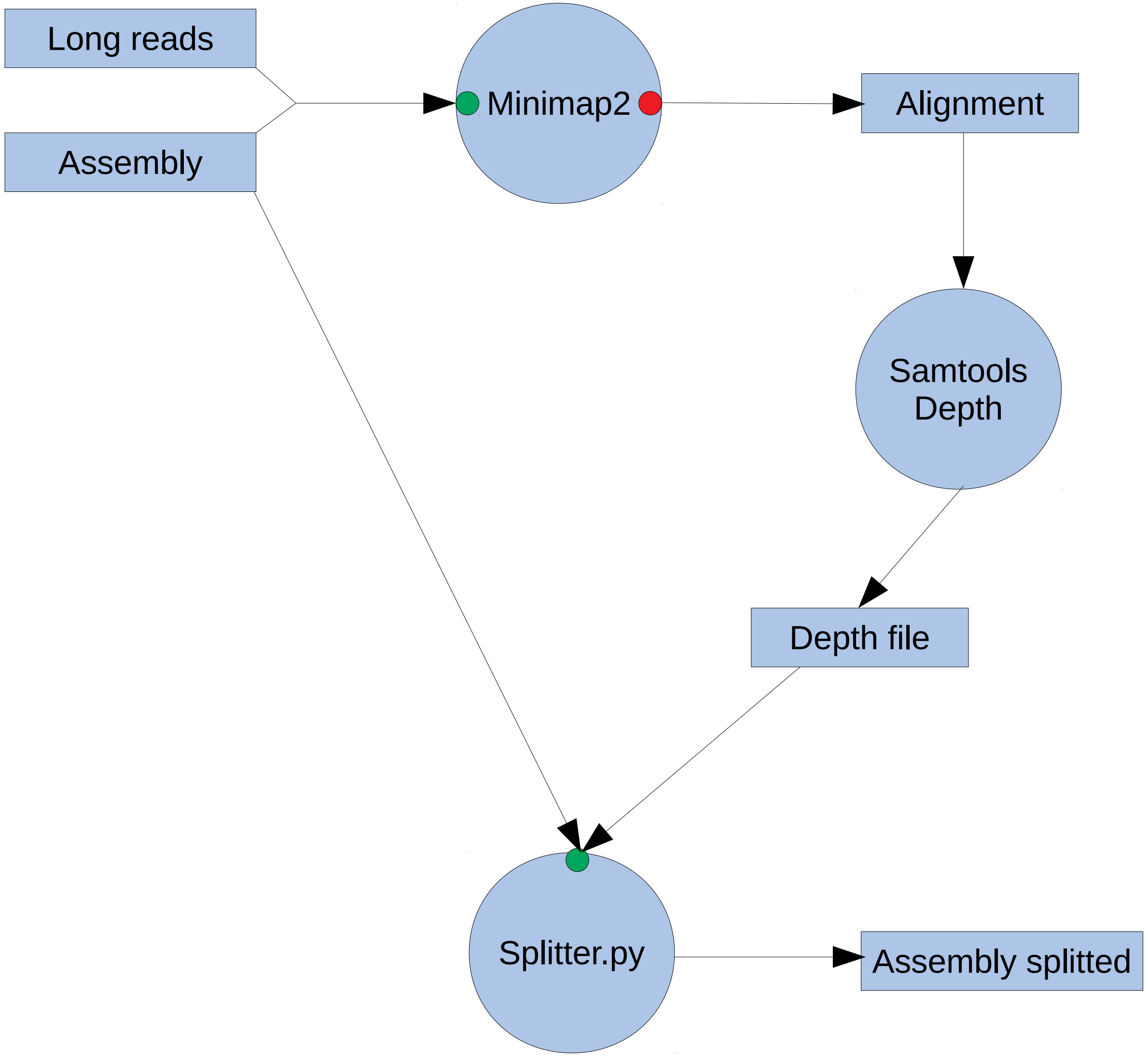
- Less than 15X long reads and 10X short reads
- Busco score from 4 % to 18 %

Running on Offspring2_37160

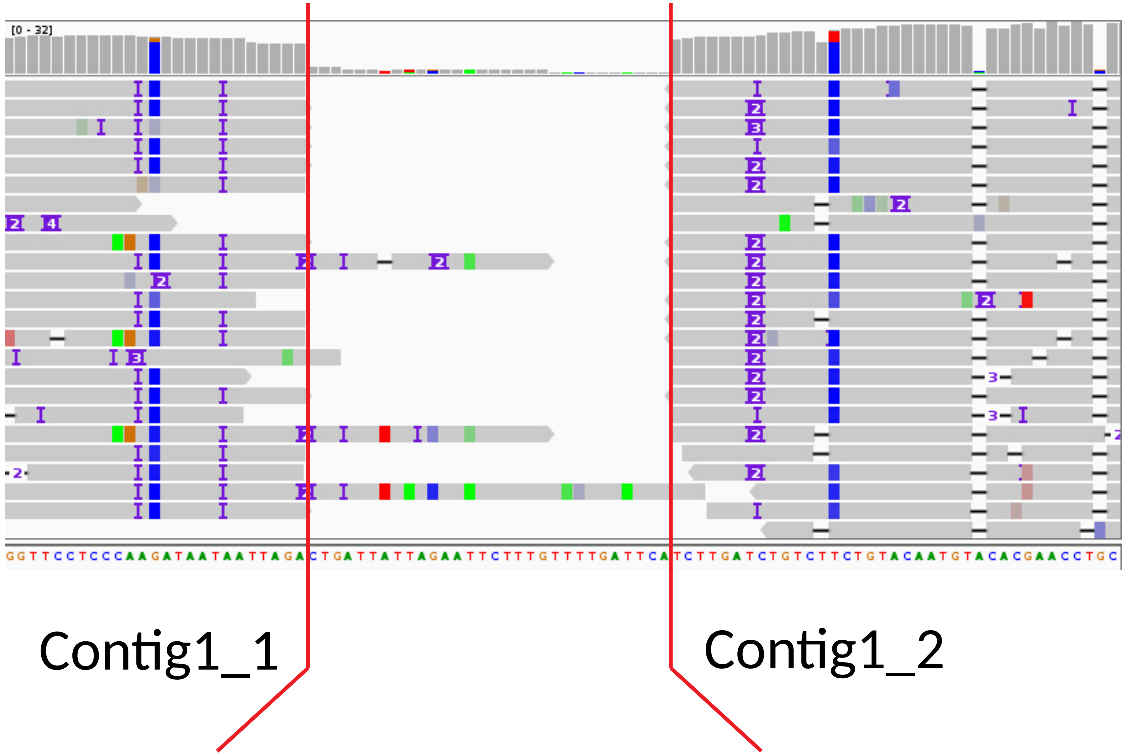
- Currently polishing with long reads
- High memory requirement
- Trying new options for long reads polishing

Already ran on fish genome

Splitting pipeline



Coverage threshold : 4X



Juicer Hi-C analysis

Huso_huso

Ameiurus_melas

Bos_taurus:

Offspring

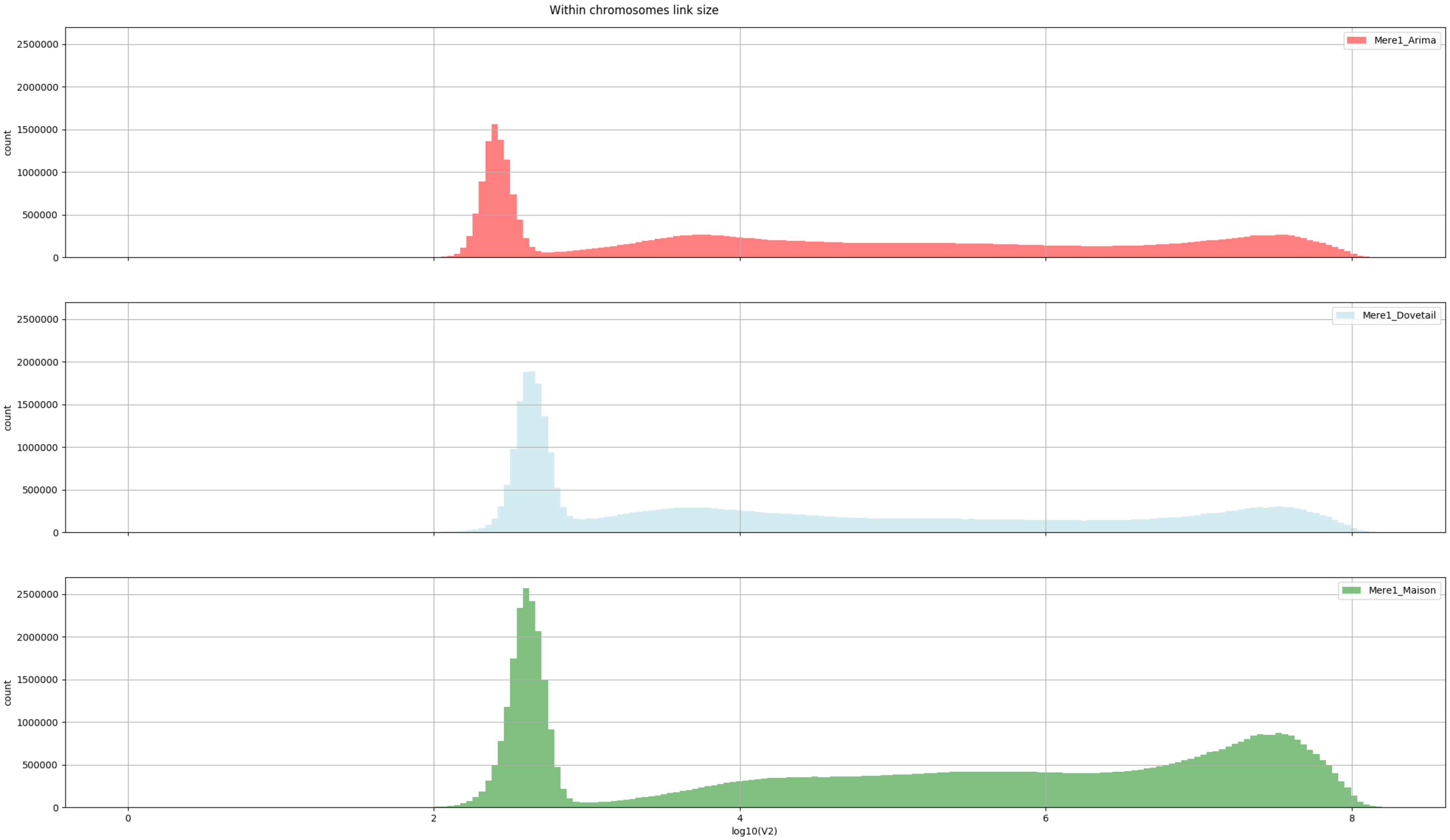
Mother

Father

Juicer Hi-C analysis

	Arima		Maison		Dovetail	
Sequenced Read Pairs:	76 522 432		134 706 615		108 635 633	
Normal Paired:	24 428 963	31.92%	75 149 462	55.79%	64 168 763	59.07%
Chimeric Paired:	34 531 814	45.13%	52 702 521	39.12%	32 067 817	29.52%
Chimeric Ambiguous:	16 968 467	22.17%	6 013 780	4.46%	11 876 366	10.93%
Unmapped:	593 188	0.78%	840 852	0.62%	522 687	0.48%
Ligation Motif Present:	35 449 898	46.33%	81 996 045	60.87%	58 469 772	53.82%
Alignable (Normal+Chimeric Paired):	58 960 777	77.05%	127 851 983	94.91%	96 236 580	88.59%
Unique Reads:	52 482 944	68.58 %	108 537 586	80.57 %	69 194 635	63.69 %
PCR Duplicates:	6 257 342	8.17 %	18 926 616	14.05 %	25 059 478	23.06 %
Optical Duplicates:	220 491	0.28 %	387 781	0.29 %	1 982 467	1.82 %
Library Complexity Estimate:	255 760 815		385 581 541		144 203 730	
Intra-fragment Reads:	426 299	0.56% / 0.81%	5 487 648	4.07% / 5.06%	294 602	0.27% / 0.43%
Below MAPQ Threshold:	14 339 601	18.74% / 27.32%	9 621 546	7.14% / 8.86%	22 165 947	20.40% / 32.03%
Hi-C Contacts:	37 717 044	49.29% / 71.87%	93 428 392	69.36% / 86.08%	46 734 086	43.02% / 67.54%
Ligation Motif Present:	10 573 286	13.82% / 20.15%	39 584 527	29.39% / 36.47%	18 375 387	16.91% / 26.56%
3' Bias (Long Range):		57% - 43%		89% - 11%		67% - 33%
Pair Type %(L-I-O-R):						
Inter-chromosomal:	8 476 814	11.08% / 16.15%	32 342 195	24.01% / 29.80%	11 913 850	10.97% / 17.22%
Intra-chromosomal:	29 240 230	38.21% / 55.71%	61 086 197	45.35% / 56.28%	34 820 236	32.05% / 50.32%
Short Range (<20Kb):	14 679 645	19.18% / 27.97%	18 377 756	13.64% / 16.93%	19 299 967	17.77% / 27.89%
Long Range (>20Kb):	14 560 489	19.03% / 27.74%	42 708 399	31.70% / 39.35%	15 519 997	14.29% / 22.43%

Juicer Hi-C analysis



In progress

- Evaluation assembly, assemblers and long reads data necessity
- Optimise polishing and splitting pipeline (faster, less memory)
- Get a genome that passed all stages of the assembly pipeline
- Evaluate the amount of reads needed for polishing
- Nextflow Hi-C data analysis pipeline
- Test « Linker », Genome phasing tool