SeqOccin

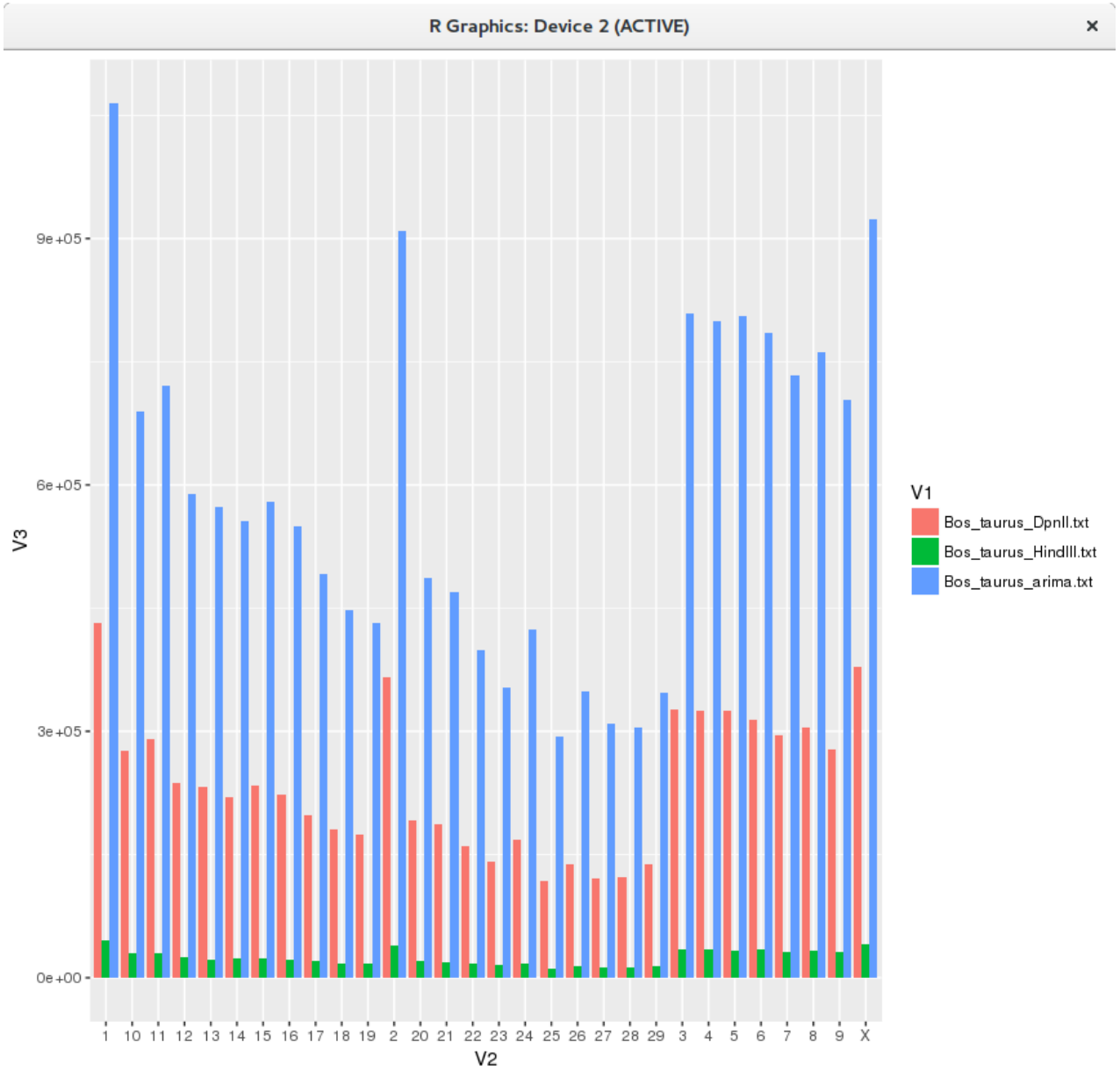
First test results on cow Hi-C Juicer processing

Christophe Klopp
http://www.sigenae.org
June 2019

Lineout

- Per chromosome restriction enzyme sitec counts
- Juicer statistics
- Hi-C links span histograms

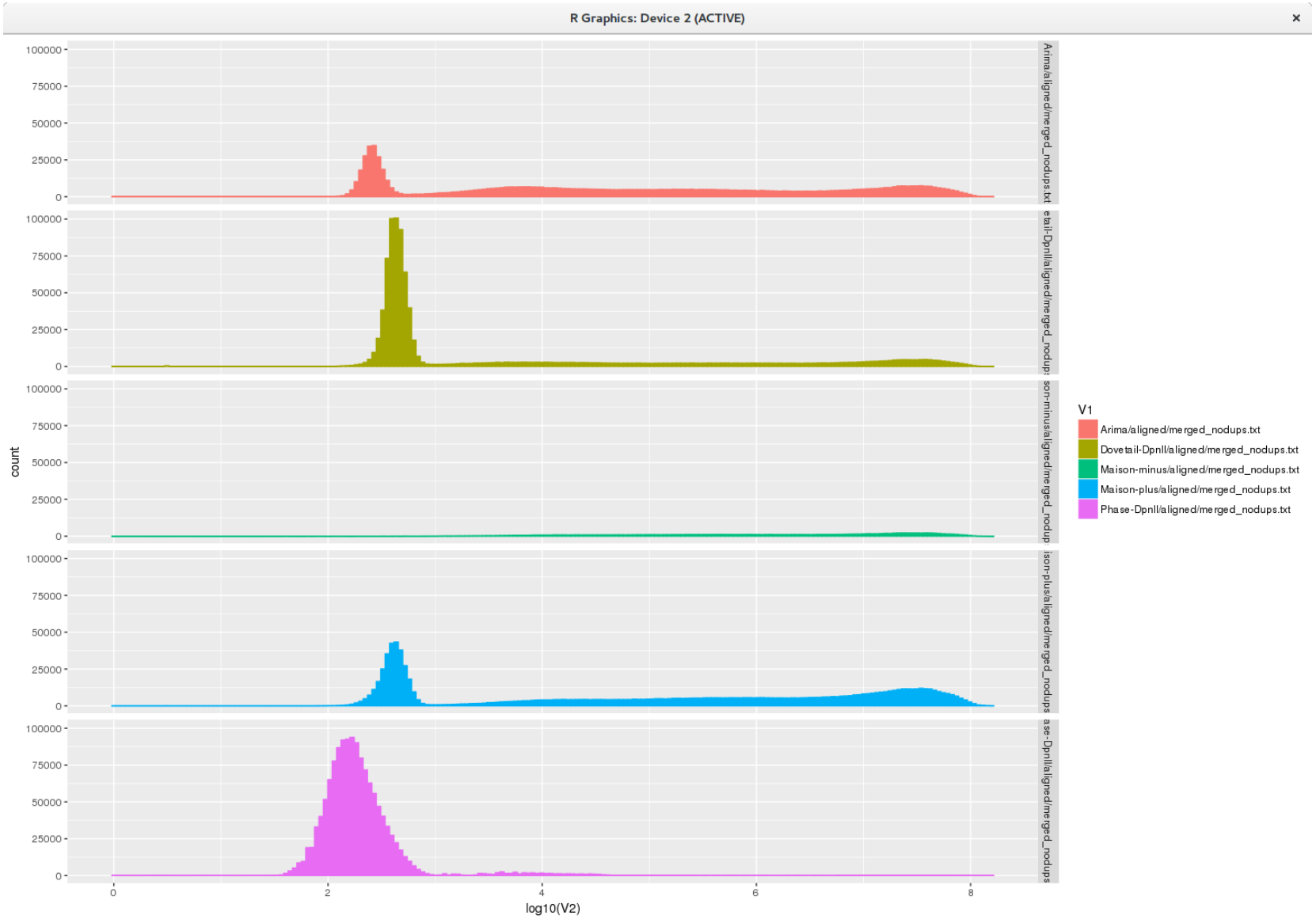
Number of restriction sites per enzyme



Juicer statistics

	Maison-plus	Maison-minus	Phase Mbol	Dovetail Mbol	Arima
Sequenced Read Pairs	1 715 720	1 541 573	1 730 569	1 496 149	1 688 139
Normal Paired	968 531	246 755	1 680 497	1 089 822	494 831
Chimeric Paired	639 571	29	7 324	264 673	841 999
Chimeric Ambiguous	76 322	1 294 788	8 596	112 476	337 716
Unmapped	31 296	0	34 152	29 178	13 593
Ligation Motif Present	1 015 734	0	629	534 121	2 098 961
Alignable Normal+Chimeric Paired	1 608 102	246 784	1 687 821	1 354 495	1 336 830
Unique Reads	1 600 212	246 343	1 623 652	1 340 649	1 333 525
PCR Duplicates	2 475	187	62 634	10 534	1 836
Optical Duplicates	5 415	254	1 535	3 312	1 469
Library Complexity Estimate	518 375 859	162 423 269	22 134 169	86 206 294	485 172 697
Intra-fragment Reads	102 134	1 272	839 484	61 487	2 200
Below MAPQ Threshold	148 946	3 803	340 609	274 588	287 695
Hi-C Contacts	1 349 132	207 041	443 559	1 004 574	1 043 630
Ligation Motif Present	577 505	176 482	301	251 505	1 058 979
3' Bias Long Range					
Pair Type % L-I-O-R					
Inter-chromosomal	51 544	82 563	6 035	196 783	260 788
Intra-chromosomal	833 692	124 478	437 524	807 791	782 842
Short Range <20Kb	273 770	14 824	435 700	578 250	363 209
Long Range >20Kb	559 922	109 654	1 824	229 521	419 629

Within chromosome link sizes



Number of sites in the merged_nodups file

condition	Existing sites	Used sites
Arima	19 947 895	2 233 930
Maison-plus	753 822	711 604
Maison-minus	753 822	345 889
Phase	9 120 443	1 496 430
Dovetail	9 120 443	2 017 366

- 2 sites per link (sequence pair)
- Close to saturation for HindIII

Arima ligation motifs

Number of motifs per read pair

```
     1       2       3       4       5       6       7       8       9      10
1107172  340364   48596   22737    8051    3734    1144     356      55      28
    11      13
     2       1
```

Number of motifs read

```
     1       2       3       4       5       6       7
1634914  176534   28342    5627     626      49       3
```

13 motifs

```
> df[df$seq == "M00185:223:000000000-CDFVM:1:1103:16221:18467", 1:3]
                                                      read pos   lmotif
166638  M00185:223:000000000-CDFVM:1:1103:16221:18467_1  30 GACTGATC
166639  M00185:223:000000000-CDFVM:1:1103:16221:18467_1  68 GACTGATC
262436  M00185:223:000000000-CDFVM:1:1103:16221:18467_2  79 GACTGATC
262437  M00185:223:000000000-CDFVM:1:1103:16221:18467_2 102 GACTGATC
868013  M00185:223:000000000-CDFVM:1:1103:16221:18467_1  76 GATCAGTC
868014  M00185:223:000000000-CDFVM:1:1103:16221:18467_1  99 GATCAGTC
969117  M00185:223:000000000-CDFVM:1:1103:16221:18467_2  22 GATCAGTC
969118  M00185:223:000000000-CDFVM:1:1103:16221:18467_2 110 GATCAGTC
1139207 M00185:223:000000000-CDFVM:1:1103:16221:18467_1  34 GATCGATC
1139208 M00185:223:000000000-CDFVM:1:1103:16221:18467_1  72 GATCGATC
1139209 M00185:223:000000000-CDFVM:1:1103:16221:18467_1 122 GATCGATC
1539716 M00185:223:000000000-CDFVM:1:1103:16221:18467_2  56 GATCGATC
1539717 M00185:223:000000000-CDFVM:1:1103:16221:18467_2 106 GATCGATC
```

M00185:223:000000000-CDFVM:1:1103:17382:5074 (11 motifs)

	read	pos	lmotif
62329	M00185:223:000000000-CDFVM:1:1103:17382:5074_2	127	GAATGATC
164662	M00185:223:000000000-CDFVM:1:1103:17382:5074_1	30	GACTGATC
164663	M00185:223:000000000-CDFVM:1:1103:17382:5074_1	91	GACTGATC
164664	M00185:223:000000000-CDFVM:1:1103:17382:5074_1	114	GACTGATC
260546	M00185:223:000000000-CDFVM:1:1103:17382:5074_2	65	GACTGATC
967151	M00185:223:000000000-CDFVM:1:1103:17382:5074_2	108	GATCAGTC
967152	M00185:223:000000000-CDFVM:1:1103:17382:5074_2	131	GATCAGTC
1056114	M00185:223:000000000-CDFVM:1:1103:17382:5074_1	95	GATCATTC
1130757	M00185:223:000000000-CDFVM:1:1103:17382:5074_1	34	GATCGATC
1130758	M00185:223:000000000-CDFVM:1:1103:17382:5074_1	118	GATCGATC
1531702	M00185:223:000000000-CDFVM:1:1103:17382:5074_2	104	GATCGATC

```
>M00185:223:000000000-CDFVM:1:1103:17382:5074 1:N:0:CTTGTA
GATCACGTAGCTACCATGCACTGGTCACGTGACTGATCGATCACGTGGCTATCATGCACT
GATAACTTGGCTGGTAATATACTGATCACATGACTGATCATTCACTGTTCACGTGACTGA
TCGATCAGCCACGTGATCAGTGCATGACCA
>M00185:223:000000000-CDFVM:1:1103:17382:5074 2:N:0:CTTGTA
GATCATGCACTGATCACATAATTGATCATGCACTGTTCACCTGGCTTATCATGCATTGAT
CATGTGACTGATCACGTGCCTGGTCATGCACTGATCACGTGGCTGATCGATCAGTCACGT
GAACAGTGAATGATCAGTCATGTGATCAGT
```

Conclusions

- Maison-minus : R2 sequencing problem which make the data unusable
- Phase : to many same site or short links
- Dovetail, Arima and Maison-plus have the same patterns
- Arima and Maison-plus give the highest counts of long range links