Short reads only *denovo* assembly of grape genomes

Michele Vidotto

IGA CLUB

July 6, 2016

UNIVERSITÀ DEGLI STUDI DI UDINE

IGA

The genome assembly problem

It's a bit like trying to do the hardest puzzle you can imagine!

Genome assembly pipeline - preprocessing

Fragment reads

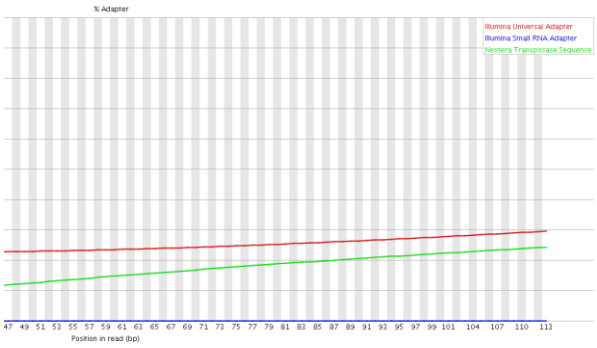
Jumping reads

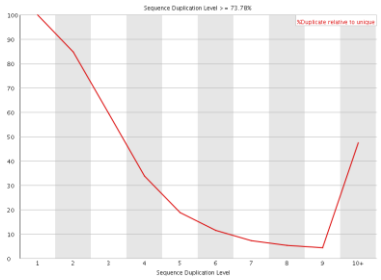
Short insert-size, paired-end reads

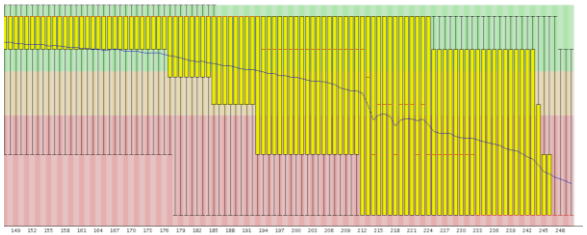
Long insert-size, paired-end reads (Mate Pair)

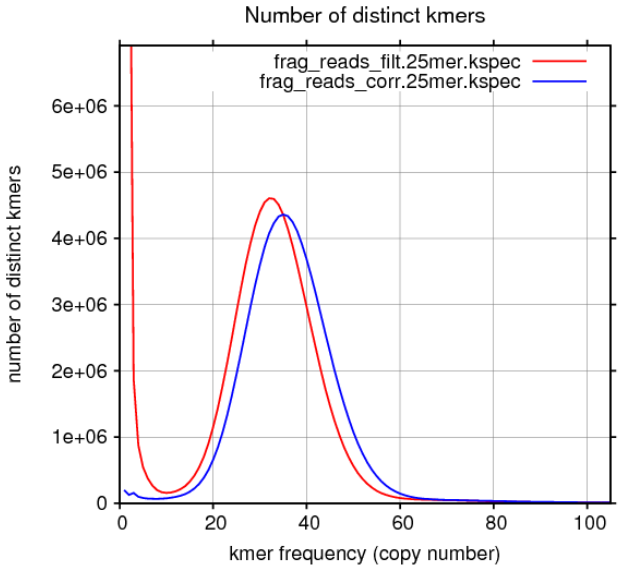
Preprocessing and filtering

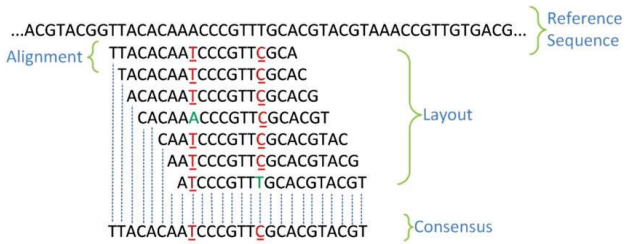
Error correction









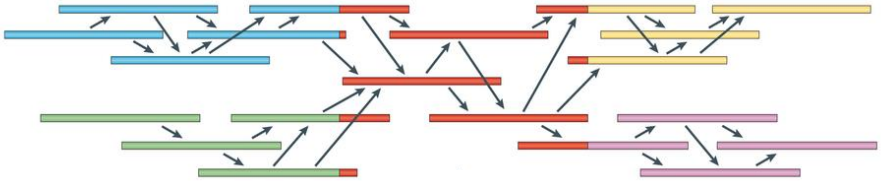


Genome assembly pipeline - graph construction

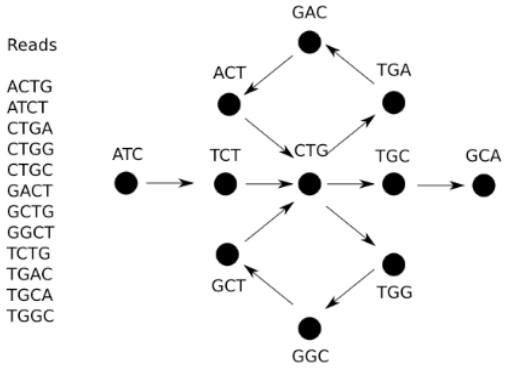
(Partially) corrected fragment reads

Graph Construction

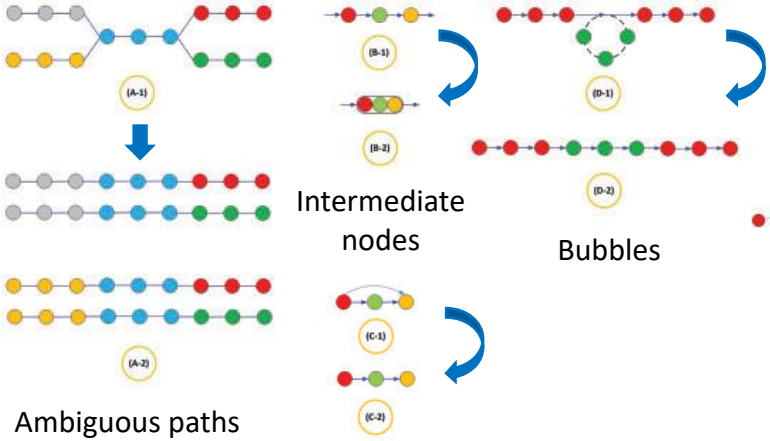
Graph Semplification

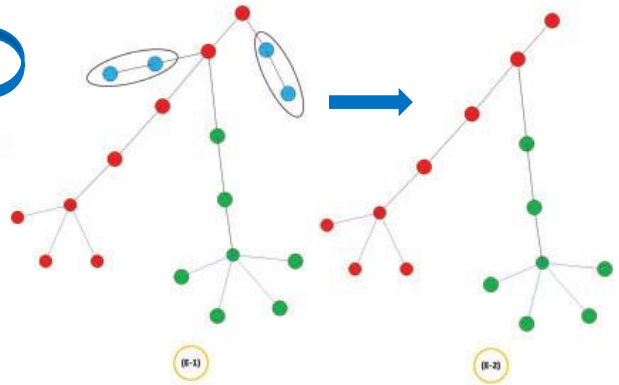


Overlap layout Consensus (OLC)



De Bruijn

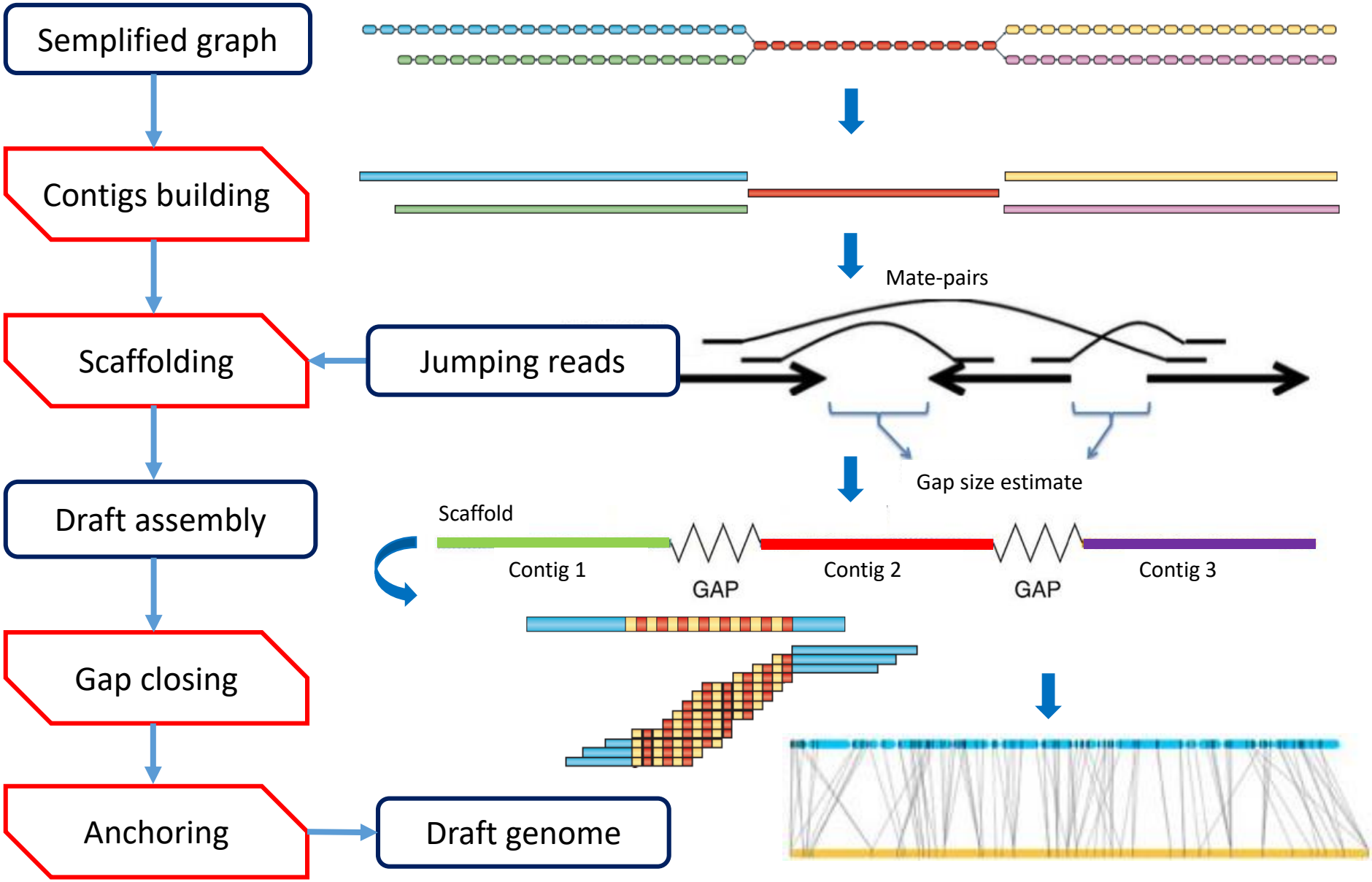




Tips (dead ends)

Michele Vidotto michele.vidotto@uniud.it

Genome assembly pipeline - postprocessing



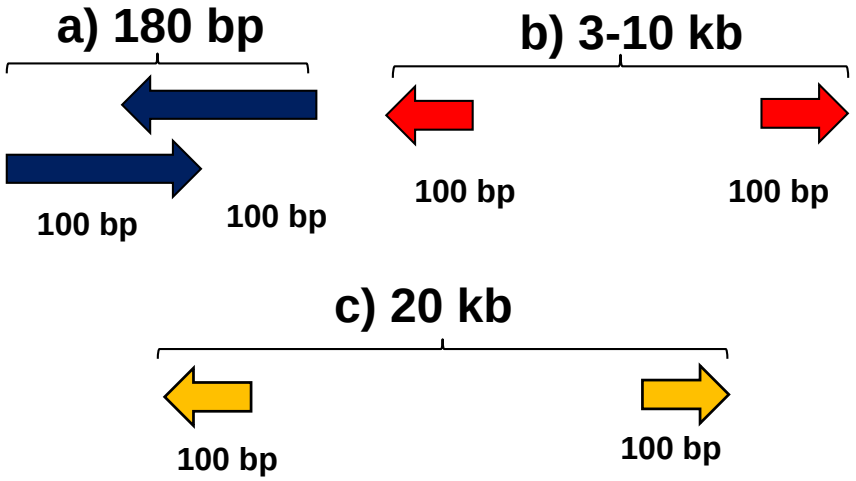
The ALLPATHS-LG assembler recipe

- **Best interaction between experimental design and algorithmic approach**

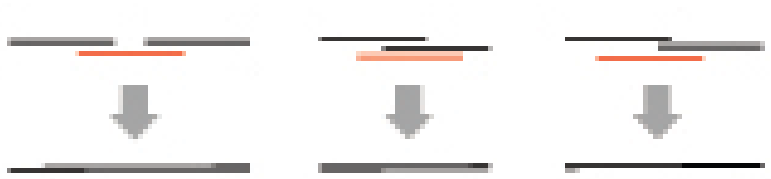
a) ~45X Overlapping fragment libraries

b) ~45X Jumping libraries

c) ~1X Long Jumping libraries

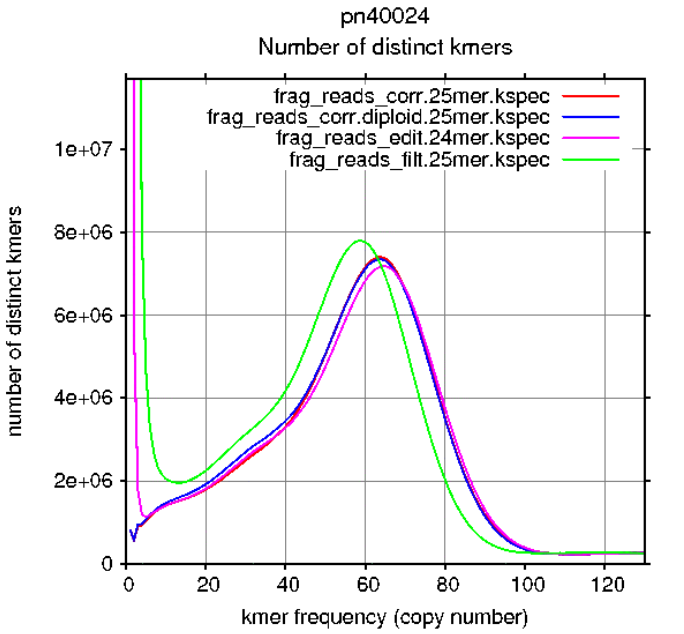


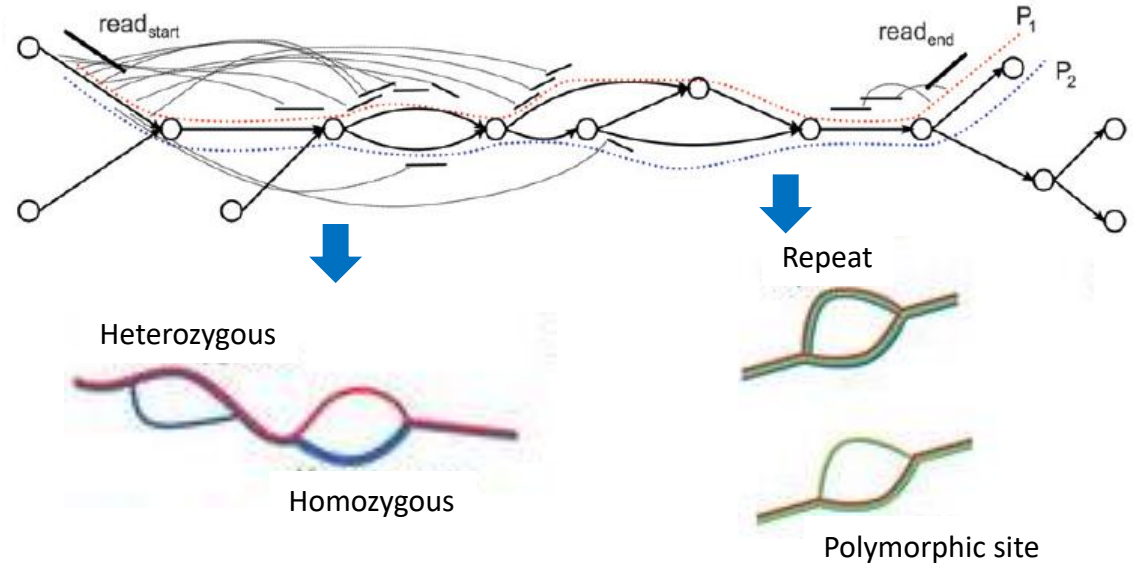
- **Debruijn graph based**
- **Overlapping paired reads merged into single long fragment**
- **Longer K-mer can be used for graph construction (k=96)**



The ALLPATHS-LG peculiarity

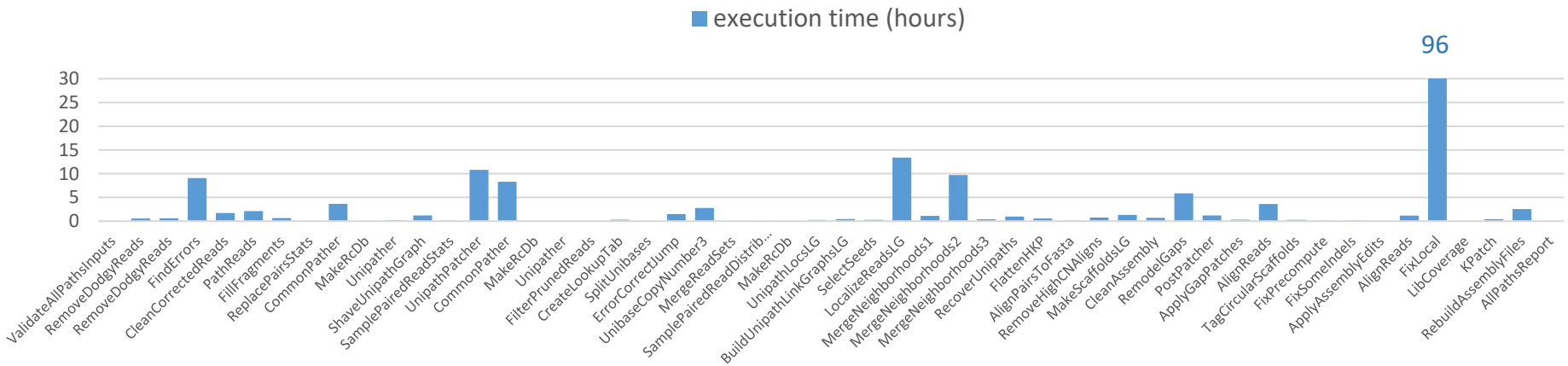
- **Multiple K-spectrum based error correction of fragment libraries (k=24,25)**
- **Alignment and error correction of the jumping reads on the assembly graph**
- **Scaffolding gap filling and modules included**
- **HAPLOIDIFY function identifies bubbles in graph with haploid:diploid copy number and resolves them by choosing one path**

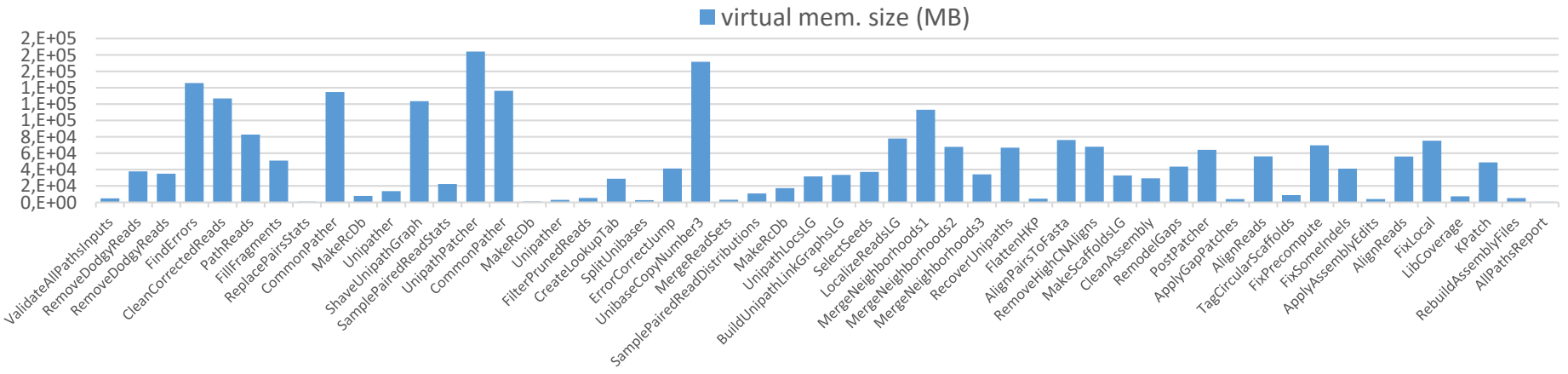




The ALLPATHS-LG pipeline

- **Modular system. Each module performing an assembly step**
- **52 modules needed for the grape assemblies**





Our ALLPATHS-LG recipe

	Cov. (X)
Traminer	
overlapping fragment	20
fragment	33
	53
jumping	19
long jumping	16
	36
Gouais Blanc	
overlapping fragment	28
fragment	35
	62
jumping	16
long jumping	14
	30
Cabernet Franc	
overlapping fragment	55
	55
jumping	23
long jumping	10
	33

	Cov. (X)
Rkatsiteli	
overlapping fragment	22
fragment	32
	55
jumping	29
long jumping	12
	40
Kishmish Vatkana	
overlapping fragment	29
fragment	23
	52
jumping	24
long jumping	10
	35
Sangiovese	
overlapping fragment	33
fragment	34
	68
jumping	23
long jumping	22
	45

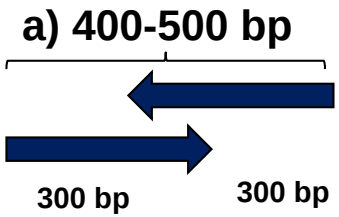
	Cov. (X)
PN40024	
overlapping fragment	11
fragment	66
	77
jumping	15
	15

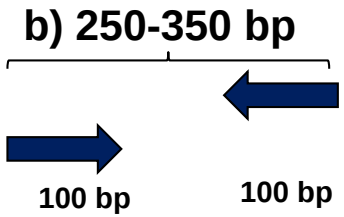
	Cov. (X)
RPV3	
overlapping fragment	29
fragment	30
	59
jumping	18
long jumping	10
	29

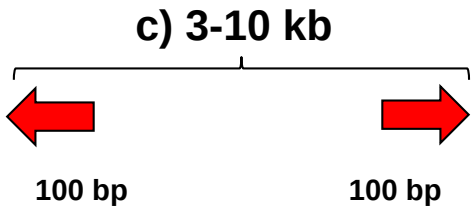
a) **~30X Overlapping fragment libraries**

b) **~20X Non-overlapping fragment libraries**

c) **~20X 3kb Jumping libraries + ~10X 10kb (long) Jumping libraries**



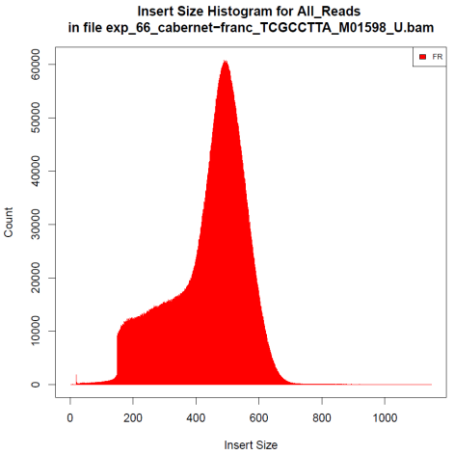


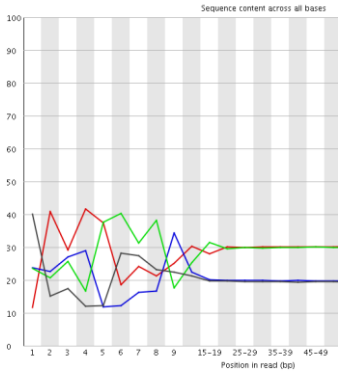


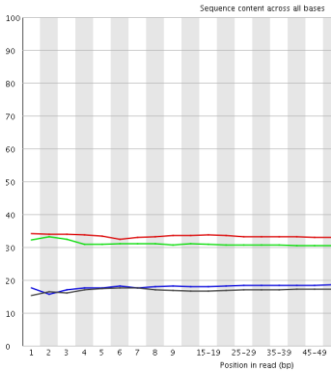
Thanks to Gabriele Magris

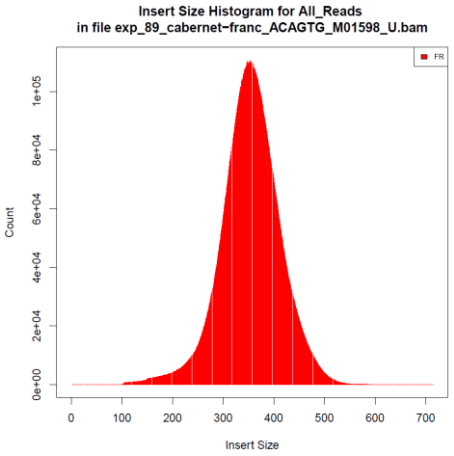
Fragment libraries – TruSeq Vs Nextera

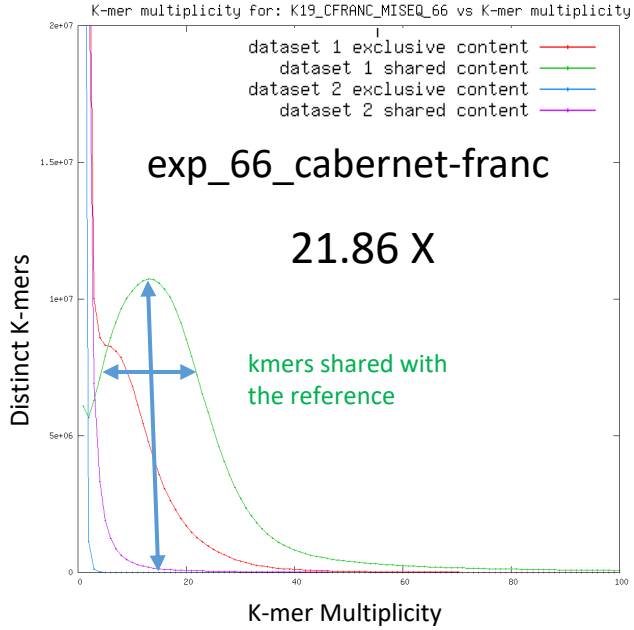
Nextera









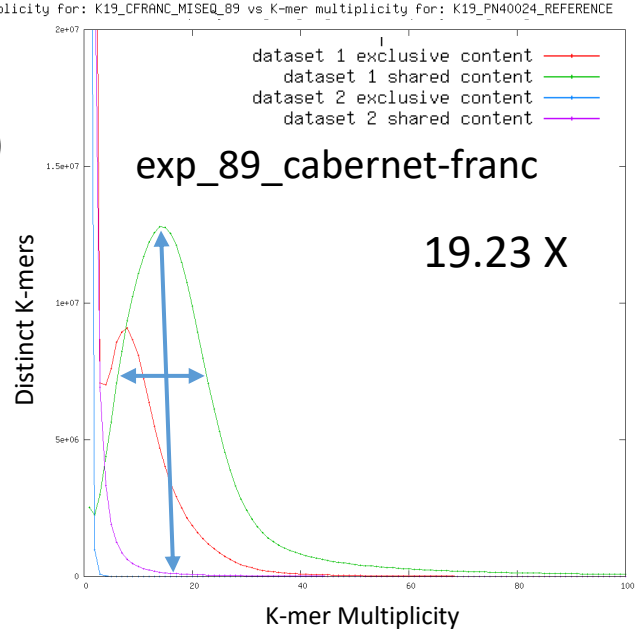


rmdup cov. BWA-mem, PN40024)

kmerisation with Jellyfish (k=19)

kmers comparison Vs PN40024
with K.A.T.

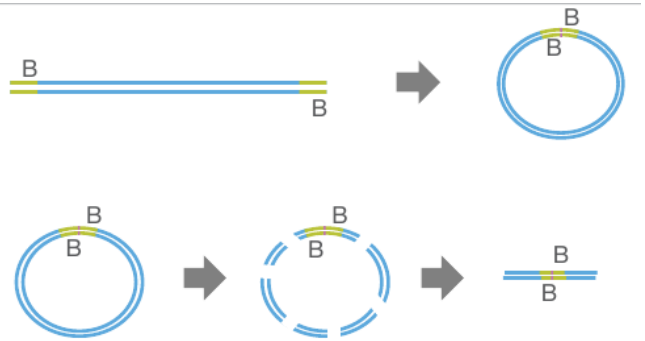
TruSeq PCR-Free

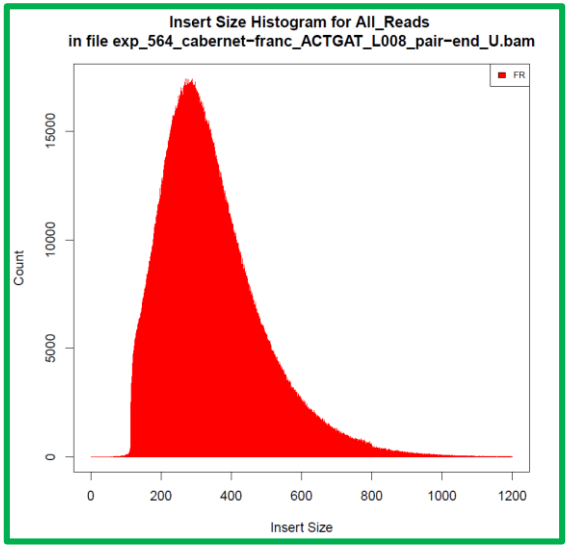


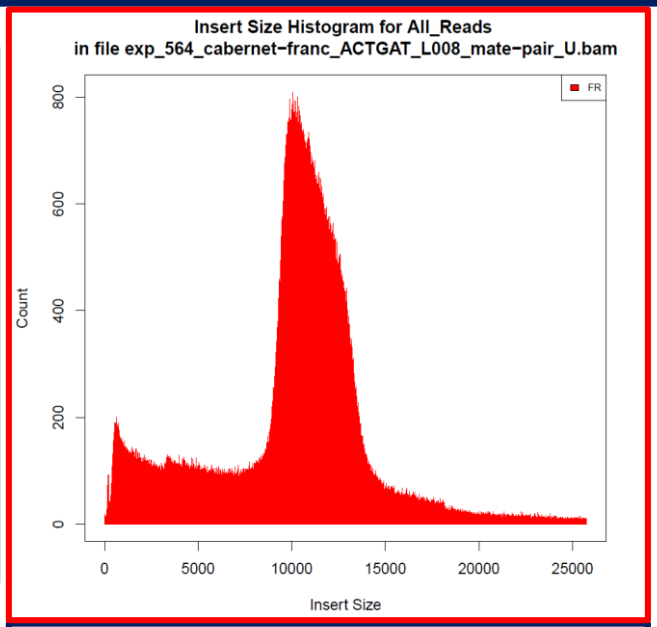
Thanks to Irena Iurman and Nicoletta Felice

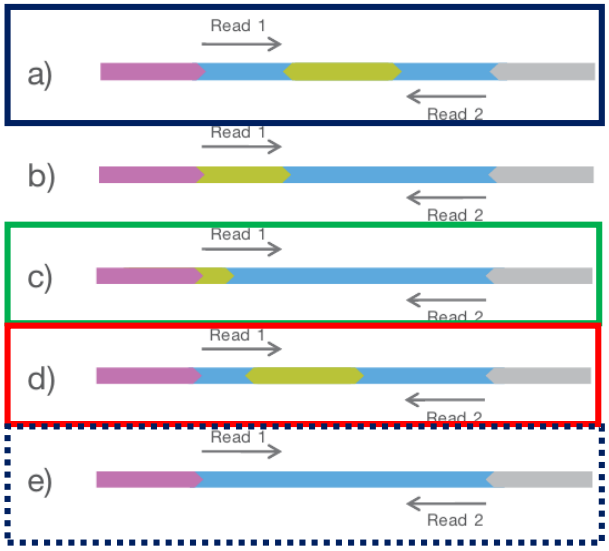
Jumping libraries – contamination problem

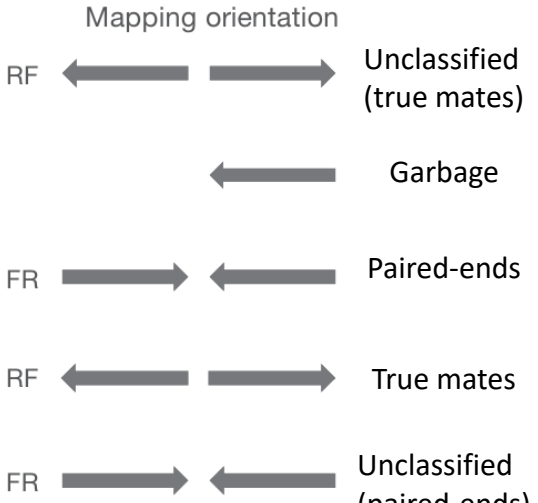
Mate-pair sequencing

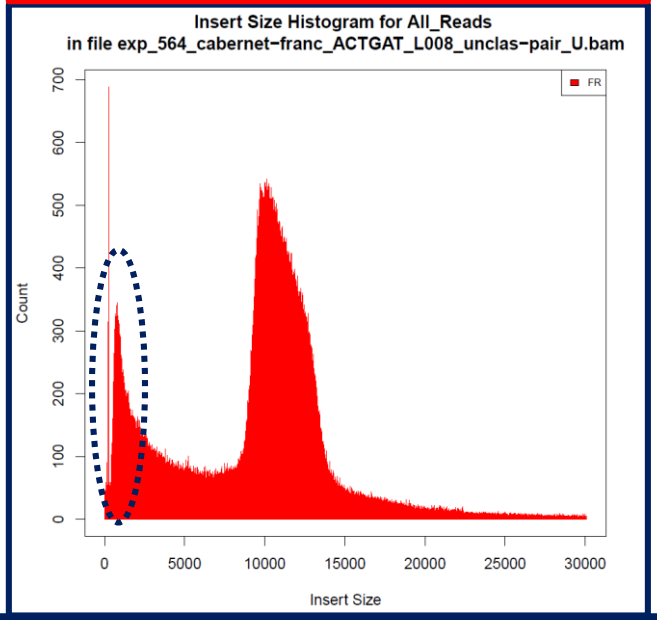






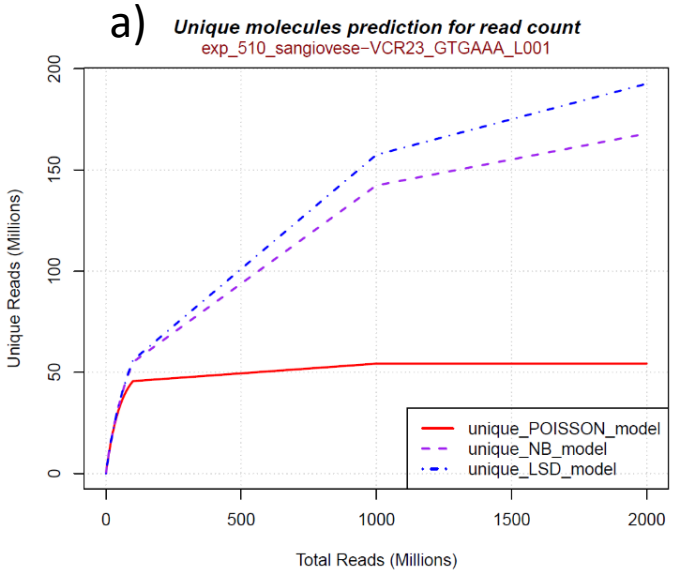




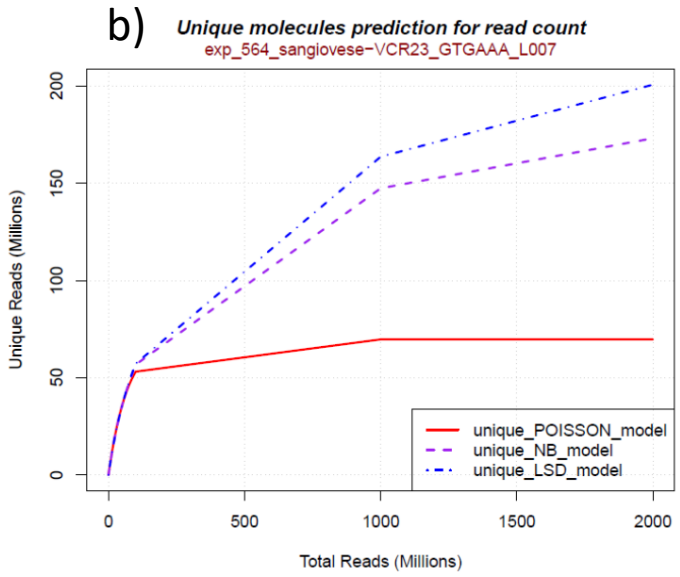


Thanks to Irena Iurman and Nicoletta Felice

Jumping libraries – duplication problem



	Unique NB model		
million reads	510	564	Δ
0	0	0	
5	4.8	4.8	-0.4%
a) 5.4	5.1		
10	9.2	9.2	-0.7%
15	13.2	13.3	-1.0%
20	16.9	17.1	-1.2%
25	20.4	20.7	-1.4%
30	23.7	24.0	-1.6%
35	26.7	27.2	-1.8%
40	29.6	30.2	-1.9%
45	32.3	33.0	-2.0%
50	34.9	35.7	-2.2%
b) 50.5		35.9	
55	37.3	38.2	-2.3%



Sangiovese_VCR23_b_library

	READS	% DUPLICATION	READ PAIR DUPLICATES	OPTICAL DUPLICATES PAIR	READS TO REACH 0.8 UNIQ FRAC (NB)	ESTIMATED LIBRARY SIZE	PRED FRAC OBSERVED (NB)
a)	5,399,709	4.80%	258,945	104,049	27,910,043	88,751,341	1.20%
b)	50,520,592	28.86%	14,581,610	5,564,168	30,297,840	96,516,581	9.60%

•**Poisson model**: the Poisson model assumes that each molecule in the library has an equal probability of being sequenced

•**Negative binomial model**: the NB model assumes there is variability in the probability of each molecule being sequenced, and that this variability follows a gamma distribution

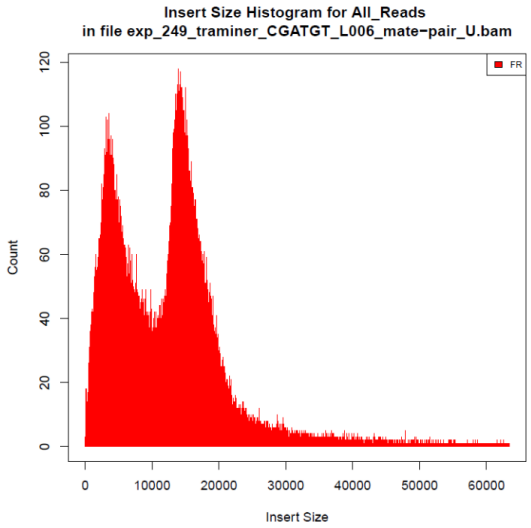
•**Log-series distribution**: the LSD is a limiting case of the NB model where the gamma distribution has infinite variance, so there are an infinite number of molecules that are infinitely unlikely to be sequenced

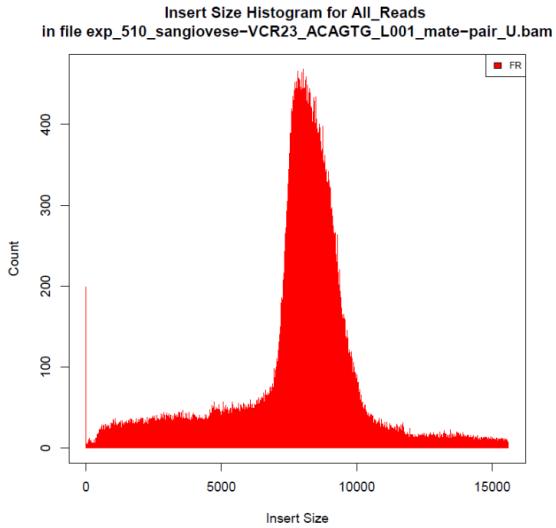
Jumping libraries – insert distribution problem

bad

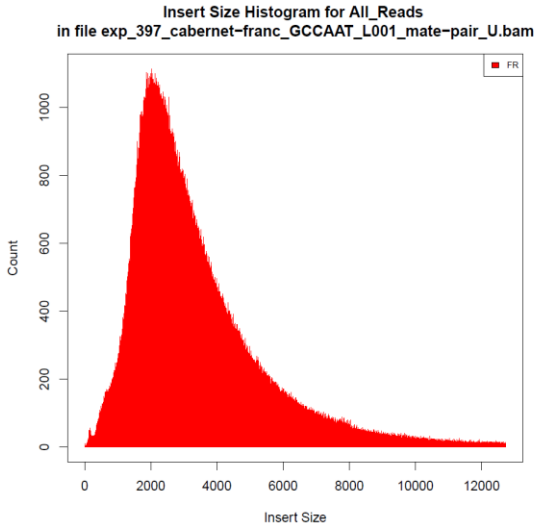
10Kb insert size

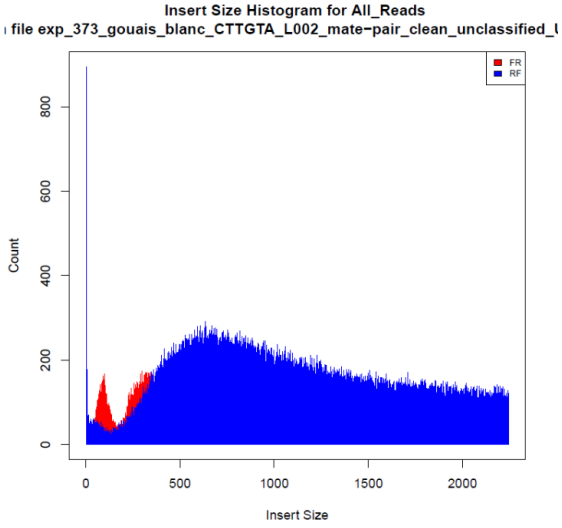
good

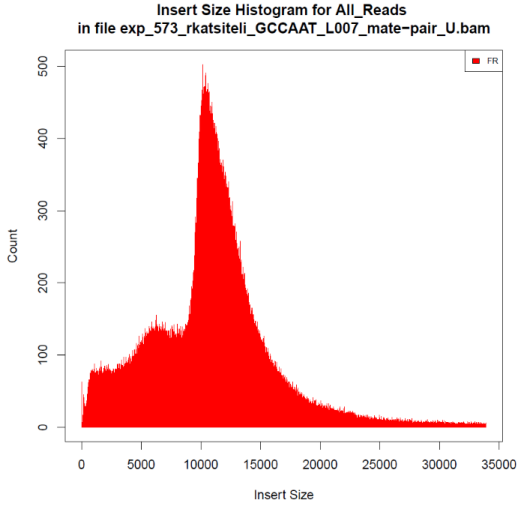


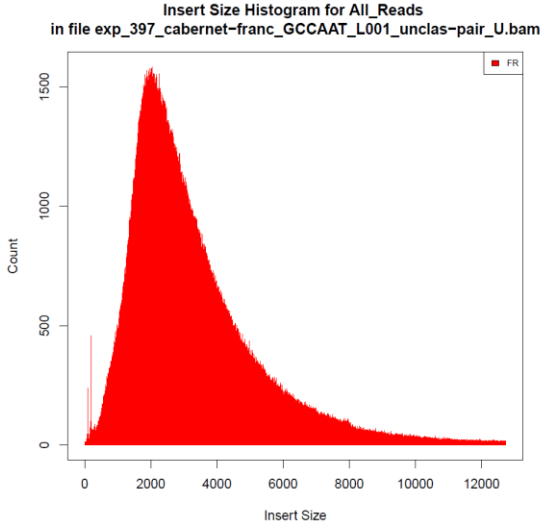


3Kb insert size
always good!







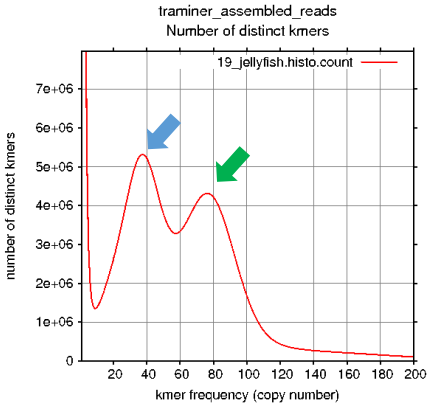


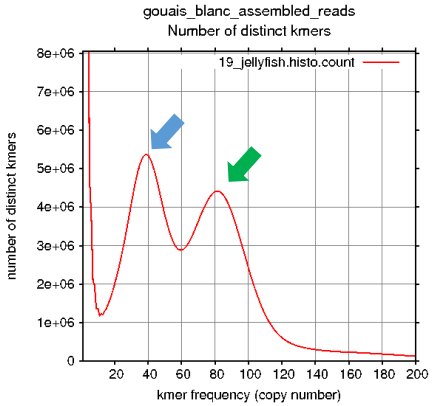
michele.vidotto@uniud.it

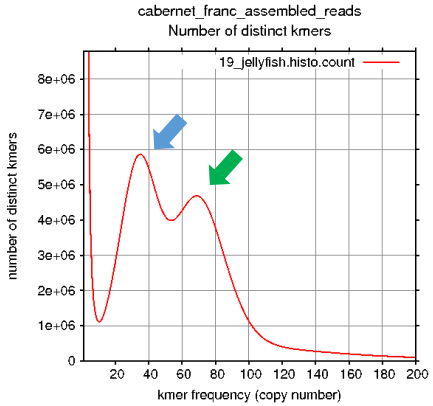
The high heterozygous grape genome

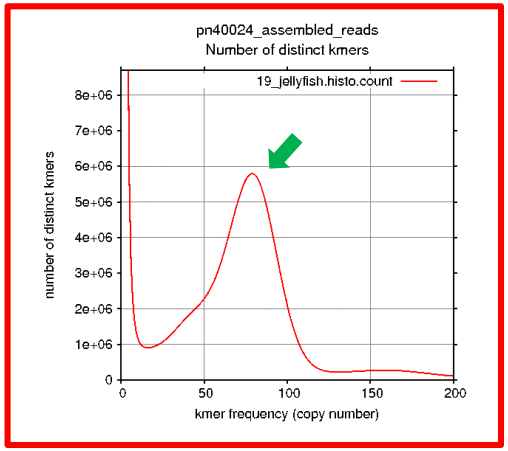
- Diploid genome
- PN4004 draft 487 Mb

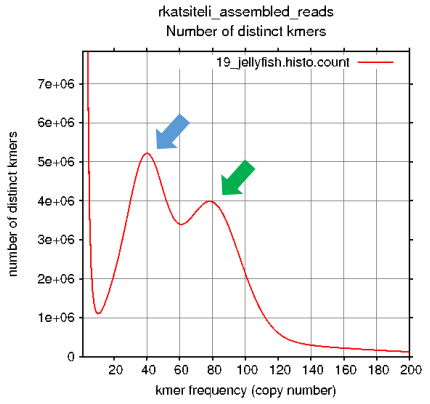
- Highly heterozygous individuals
- 19 chr

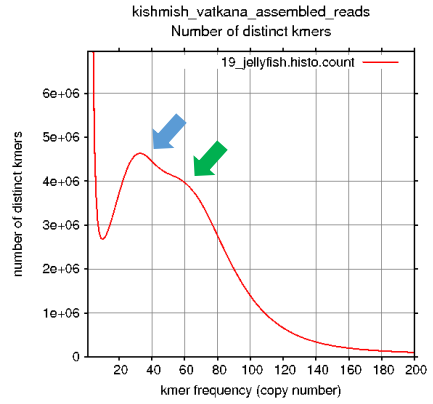


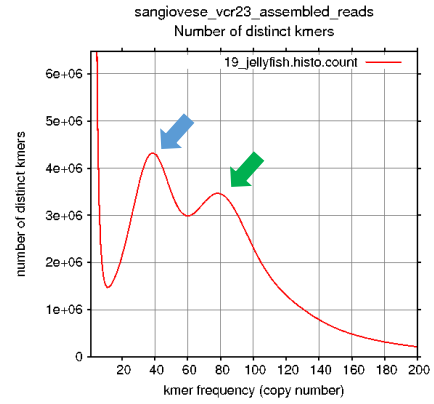


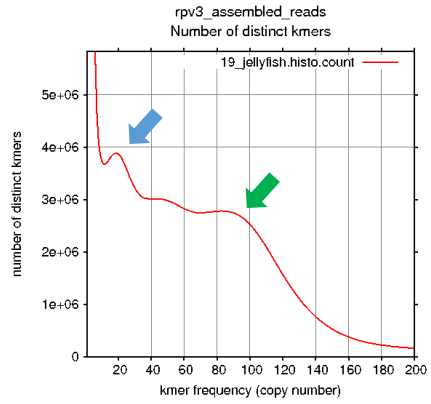












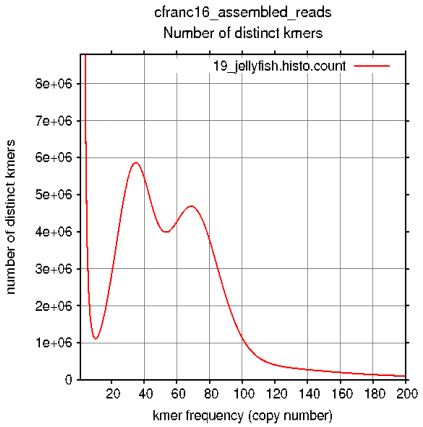
michele.vidotto@uniud.it

Assemblies results

	traminer		gouais blanc		cabernet franc		pn40024	
Genome size estimated (bp)	394,441,994		420,689,048		483,072,260		444,716,040	
Estimated CN=1 (bp)	71.1%		69.6%		59.6%		65.0 %	
Estimated CN>1 (bp)	28.9%		30.4%		40.4%		35.0 %	
Coverage estimated (X)	53		58		50		64	
	contigs	scaffolds	contigs	scaffolds	contigs	scaffolds	contigs	scaffolds
Total (bp)	412,472,184	678,662,631	415,882,525	743,730,069	487,795,456	711,707,043	377,475,581	521,491,794
Average (bp)	3,728	78,223	3,722	77,764	4,770	57,005	6,814	86,684
Max (bp)	137,674	2,817,510	113,144	2,522,649	180,444	1,496,800	148,414	1,727,814
Min (bp)	32	885	69	905	58	894	118	885
Sequences (#)	110,653	8,676	111,742	9,564	102,258	12,485	55,399	6,016
N50 (#)	12,360	484	12,724	524	9,708	1,178	7,151	553
L50 (bp)	7,841	413,025	7,541	408,071	11,145	178,391	14,075	264,099
Gaps (#)		67,484		66,217		55,190		40,378
Median gap length (bp)		4,012		5,545		4,405		3,825

	rkatsiteli		kishmish vatkana		sangiovese		rpv3	
Genome size estimated (bp)	407,302,602		454,927,661		420,525,421		435,235,299	
Estimated CN=1 (bp)	69.1 %		56.5 %		68.4 %		58.6 %	
Estimated CN>1 (bp)	30.9 %		43.5 %		31.6 %		41.4 %	
Coverage estimated (X)	52		46		64		54	
	contigs	scaffolds	contigs	scaffolds	contigs	scaffolds	contigs	scaffolds
Total (bp)	426,363,529	753,744,202	410,415,417	610,410,426	412,308,934	691,235,812	397,904,367	608,826,013
Average (bp)	3,739	74,710	3,803	71,426	4,044	44,121	3,560	69,240
Max (bp)	135,733	2,436,454	119,547	2,355,113	184,133	2,528,353	90,275	2,693,357
Min (bp)	46	889	27	895	76	881	46	894
Sequences (#)	114,026	10,089	107,916	8,546	101,962	15,667	111,757	8,793
N50 (#)	14,103	612	12,979	457	10,243	627	15,087	410
L50 (bp)	7,250	352,572	7,751	364,852	8,975	301,183	7,032	425,483
Gaps (#)		69,862		64,039		61,417		76,094
Median gap length (bp)		5,200		5,728		3,581		1,770

Assemblies evaluation – kmers comparison

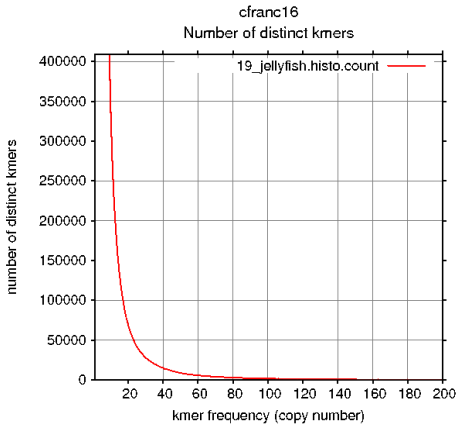


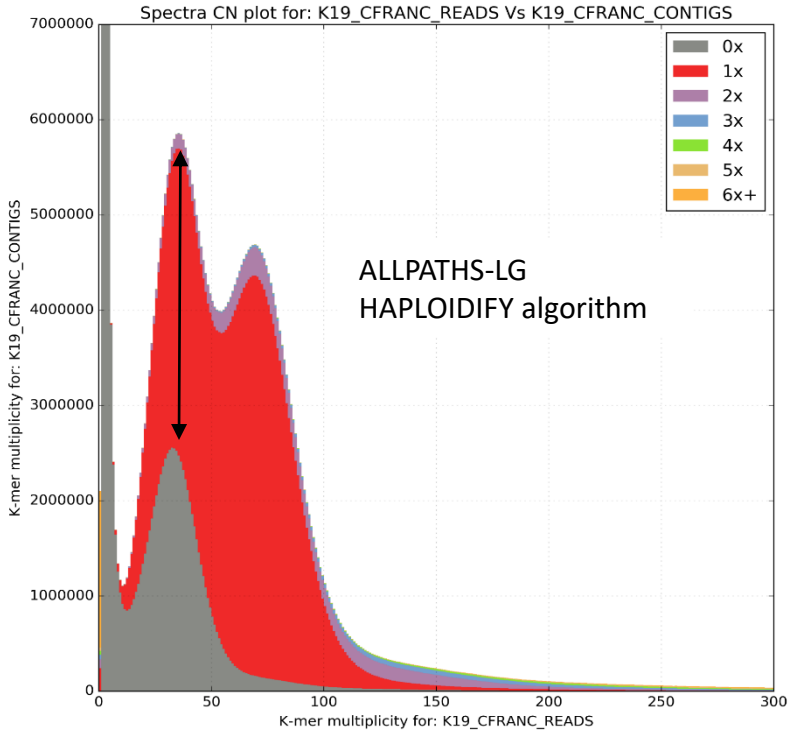
kmerisation with Jellyfish (k=19)

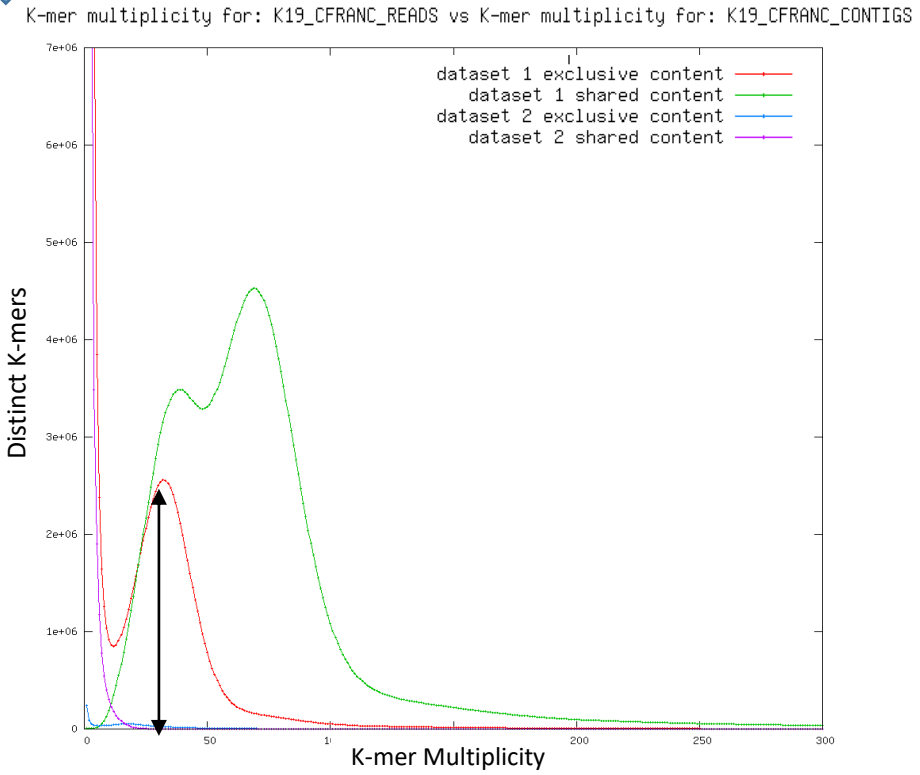
kmers comparison
reads Vs assembly with K.A.T.

fragment reads assembled

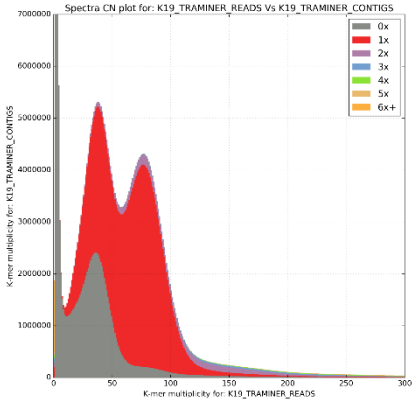
assembly contigs



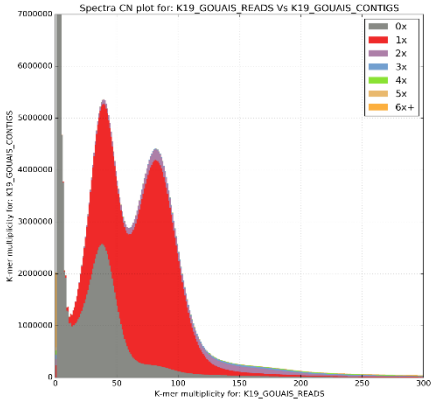




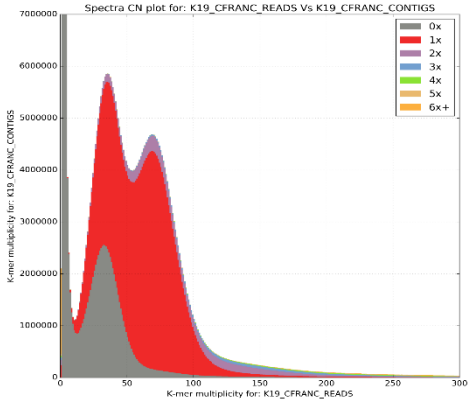
Assemblies evaluation – kmers comparison



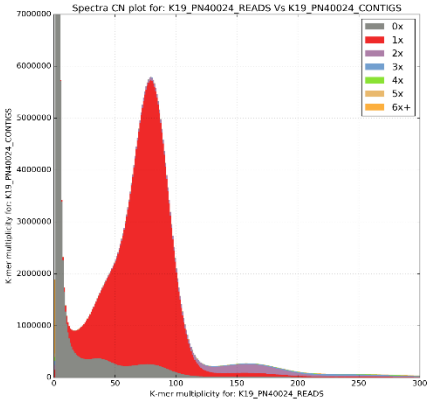
Traminer



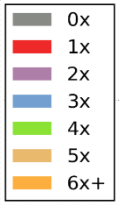
Gouais Blanc

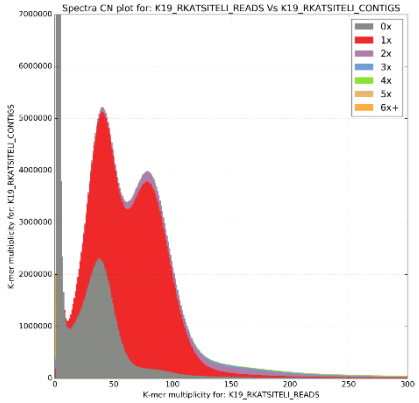


Cabernet Franc

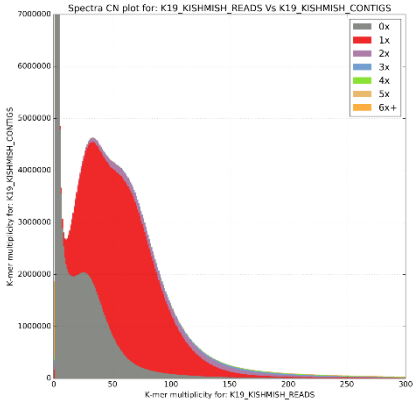


PN40024

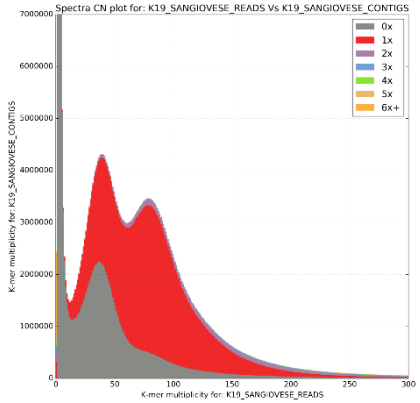




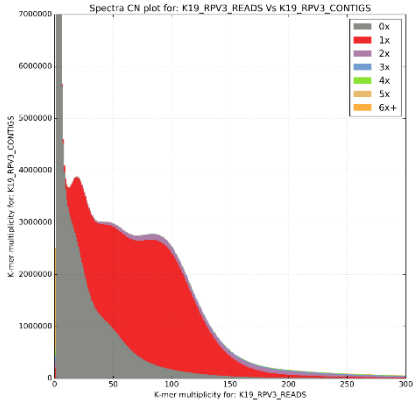
Rkatsiteli



Kishmish Vatkana

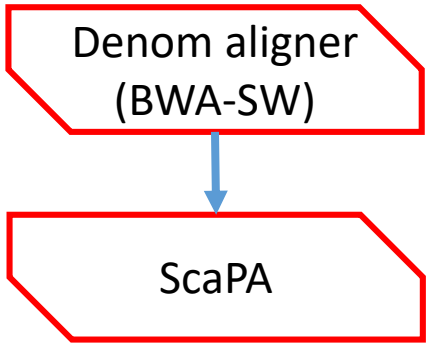


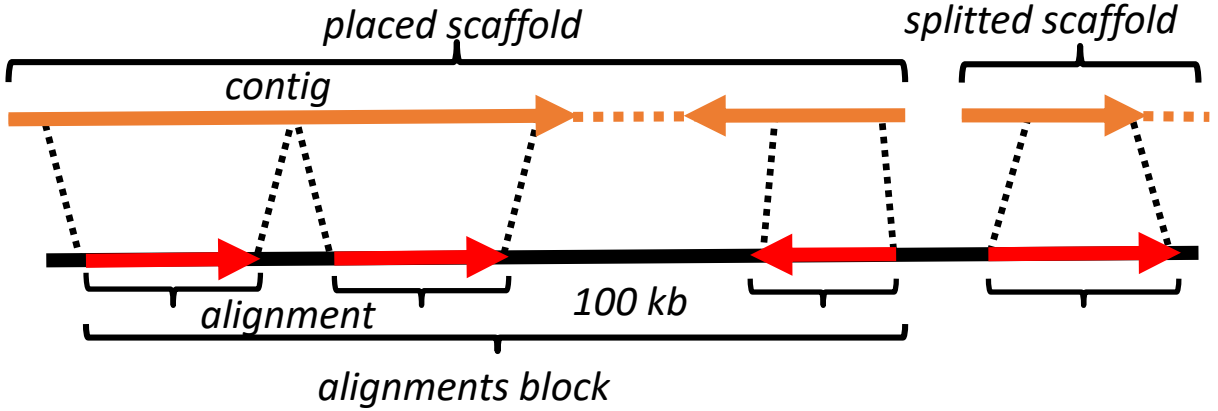
Sangiovese



Rpv3

Assemblies evaluation – alignment on the reference

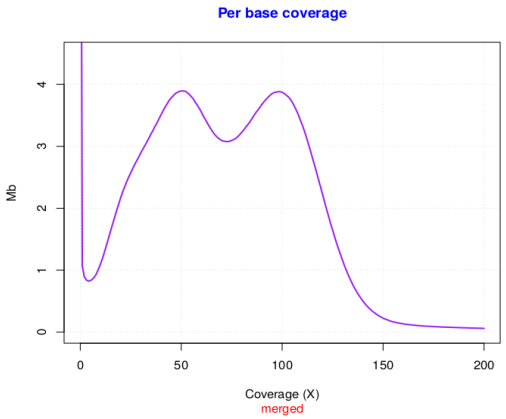




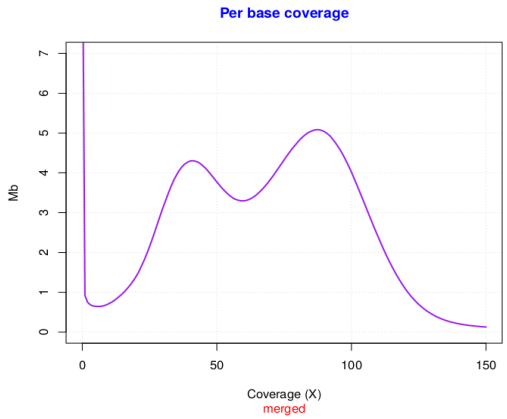
	traminer	gouais blanc	cabernet franc	pn40024	rkatsiteli	kishmish vatkana	sangiovese	rpv3
placed scaffolds:	78.96%	80.44%	83.14%	94.28%	81.15%	76.42%	81.32%	77.27%
placed bp:	87.36%	87.31%	86.21%	92.70%	85.36%	83.20%	85.64%	82.76%
anchored bp:	77.58%	77.21%	77.14%	88.06%	74.58%	73.79%	75.05%	72.69%
splitted scaffolds:	3.86%	2.77%	4.21%	3.13%	3.49%	5.21%	3.68%	5.38%
rescued bp:	5.30%	4.40%	4.30%	5.65%	5.48%	7.91%	4.79%	6.85%
unplaced scaffolds:	17.17%	16.79%	12.65%	2.59%	15.36%	18.37%	15.01%	17.35%
unplaced bp:	6.08%	7.23%	8.20%	0.77%	7.86%	7.43%	8.31%	8.57%
total placed bp:	92.66%	91.70%	90.51%	98.35%	90.84%	91.11%	90.43%	89.61%

	traminer	gouais blanc	cabernet franc	pn40024	rkatsiteli	kishmish vatkana	sangiovese	rpv3
genes:	86.61%	85.96%	88.17%	86.88%	86.07%	84.16%	85.47%	85.45%
exons:	93.52%	92.89%	94.76%	94.43%	92.68%	91.91%	93.14%	92.62%
repeats:	81.74%	80.38%	83.88%	79.66%	82.16%	79.17%	79.89%	79.95%
lnc antisense:	90.10%	89.85%	91.08%	92.79%	89.73%	86.92%	89.61%	88.88%
lnc intergenic:	81.60%	79.69%	83.11%	82.45%	81.35%	76.02%	78.28%	79.14%
lnc intronic:	92.02%	90.30%	93.73%	93.35%	91.06%	90.68%	90.68%	88.78%

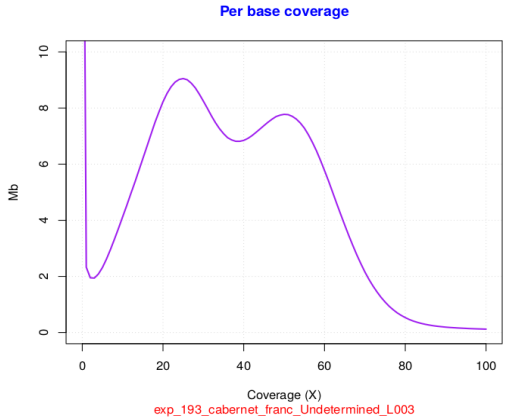
Assemblies evaluation – fragment reads relignment



Traminer

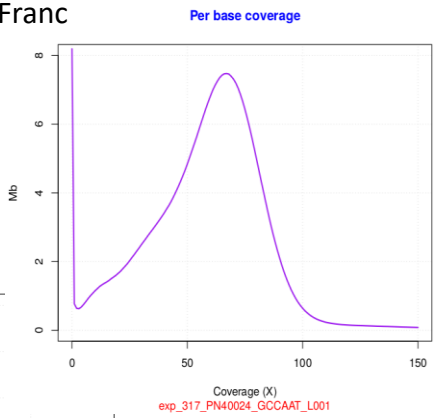


Gouais Blanc

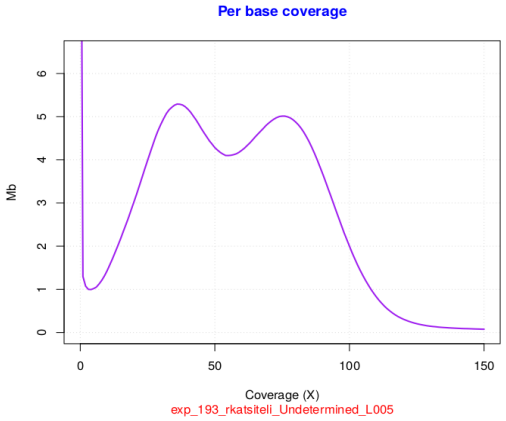


Cabernet Franc

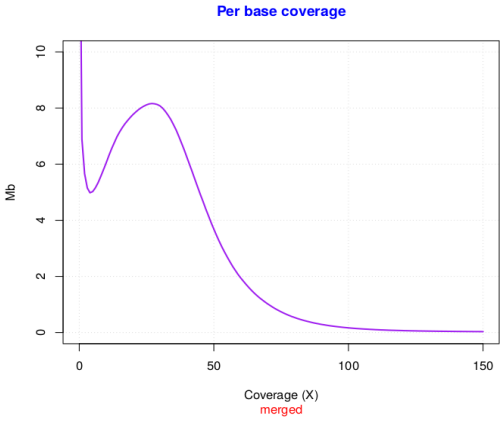
	Traminer	Gouais Blanc	Cabernet Franc	Rkatsiteli	Kishmish Vatkana	Sangiovese	PN40024
Proper pairs (%)	94.05	93.76	96.45	95.47	94.15	89.33	94.03



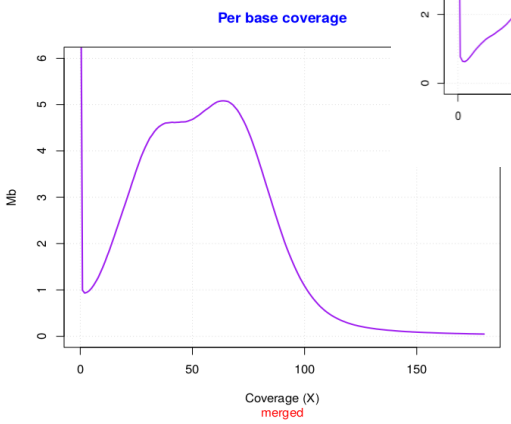
PN40024



Rkatsiteli

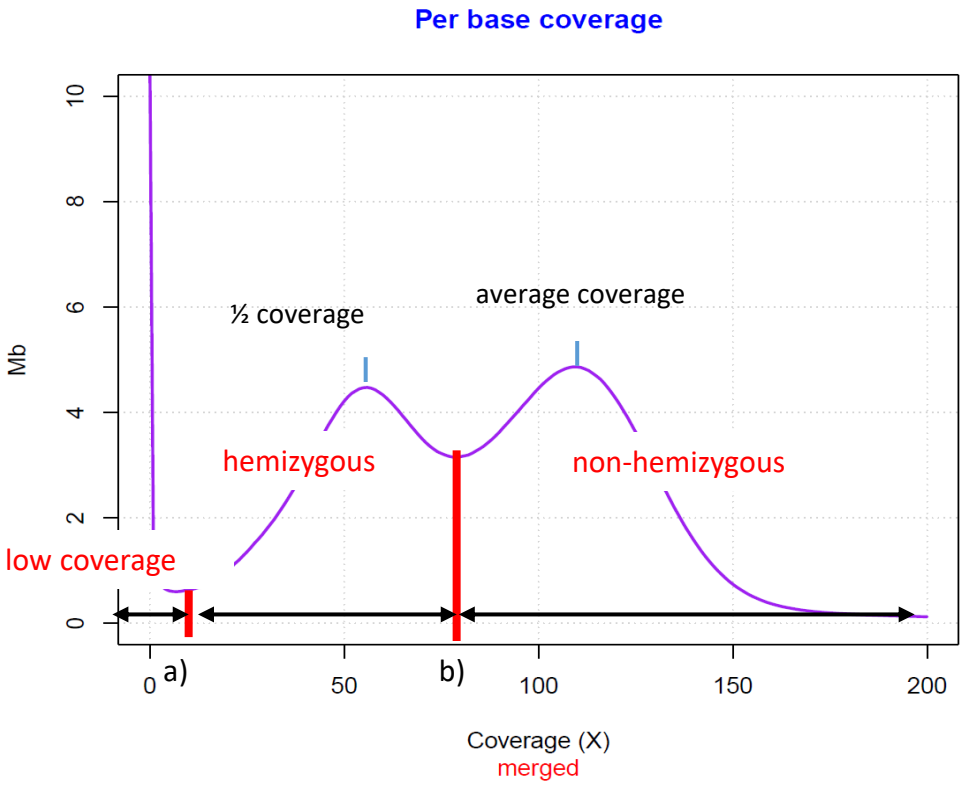


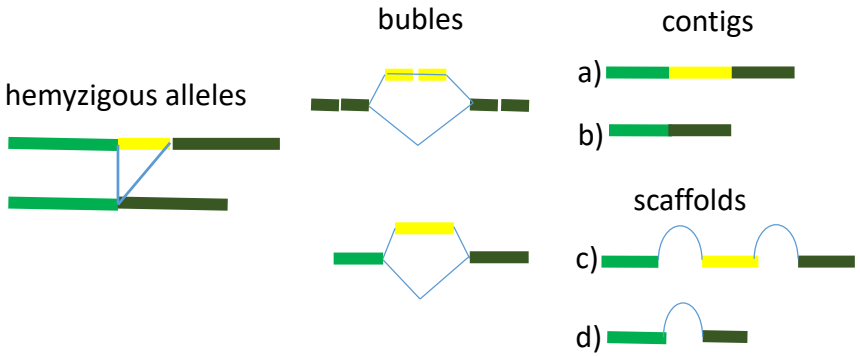
Kishmish Vatkana



Sangiovese

High emizygosity emerges

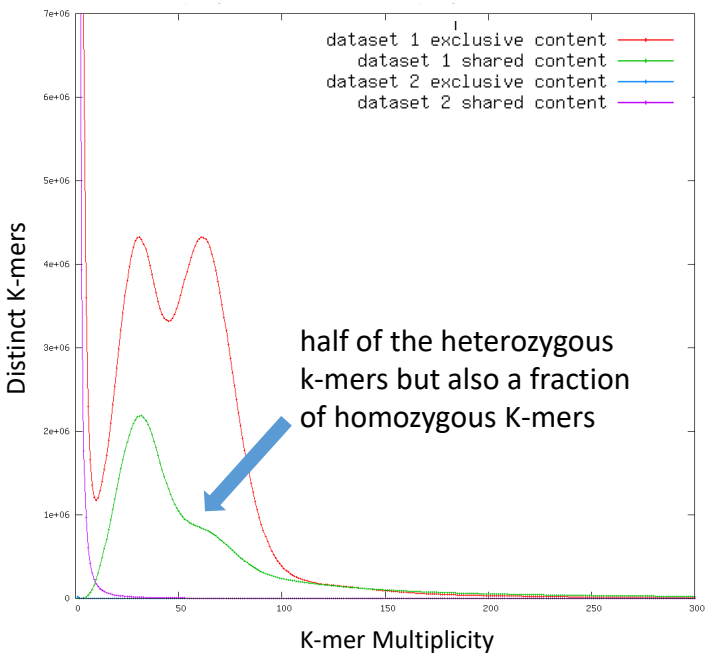




	Traminer		Gouais Blanc		Cabernet Franc		Rkatsiteli		Sangiovese	
	% of contigs	% total length	% of contigs	% total length	% of contigs	% total length	% of contigs	% total length	% of contigs	% total length
hemizygous	63.55%	40.36%	49.21%	27.46%	67.81%	41.46%	60.32%	35.95%	39.75%	20.62%
non-hemizygous	36.24%	59.55%	49.44%	72.14%	31.97%	58.46%	39.37%	63.91%	60.13%	79.33%
low coverage	0.06%	0.03%	1.26%	0.37%	0.08%	0.02%	0.15%	0.07%	0.04%	0.01%
uncovered	0.14%	0.06%	0.09%	0.03%	0.15%	0.05%	0.16%	0.07%	0.09%	0.03%

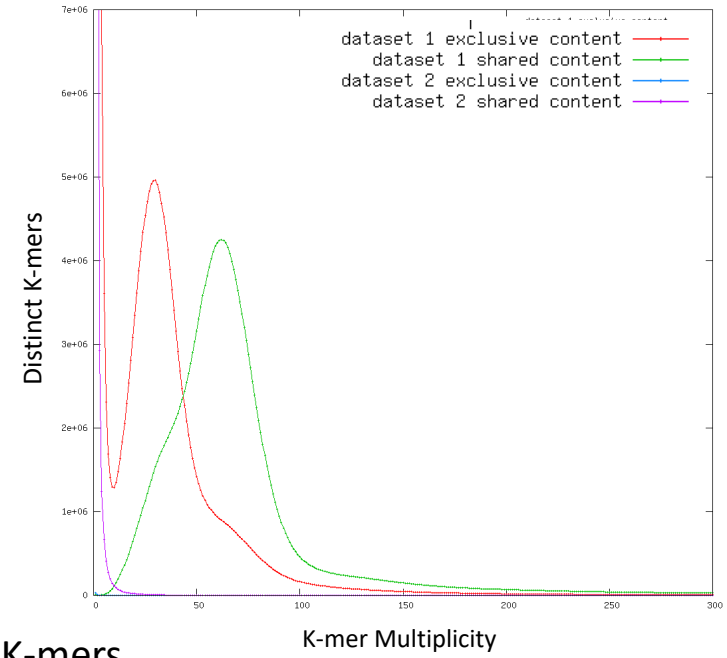
Emizygosity contigs are heterozygous and repetitive

putative hemyzygous contigs Vs fragment reads



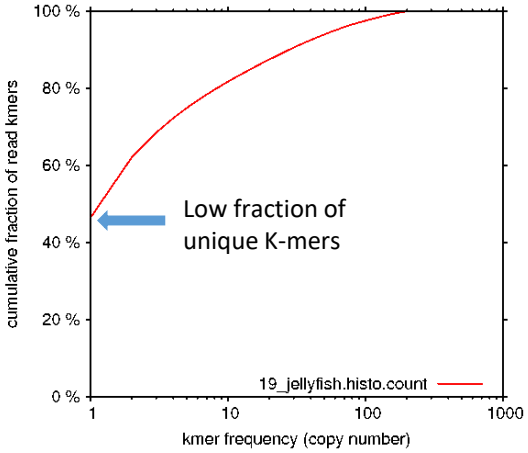
Cabernet Franc

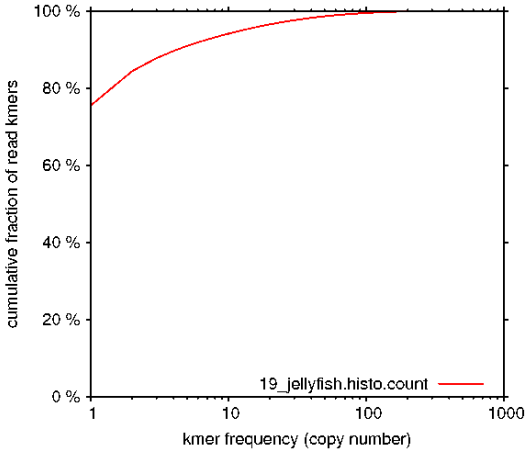
non-hemyzygous contigs Vs fragment reads



Cumulative fraction of K-mers

putative hemyzygous contigs





putative non-hemyzygous contigs

Michele Vidotto
michele.vidotto@uniud.it

Aknowledgment

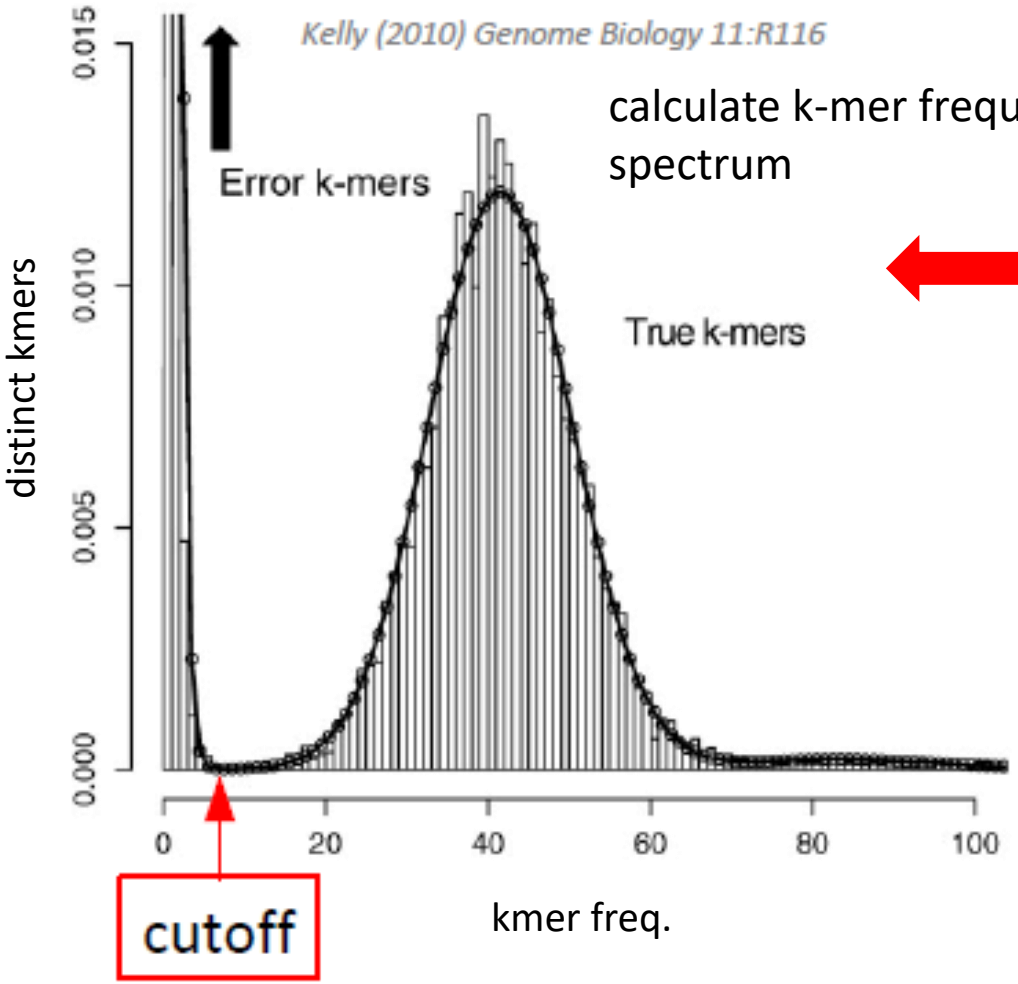
Prof. Michele Morgante
Fabio Marroni
Gabriele Magris
Sara Pinosio

IGATS
Davide Scaglione
Simone Scalabrin
Irena Iurman
Vittorio Zamboni

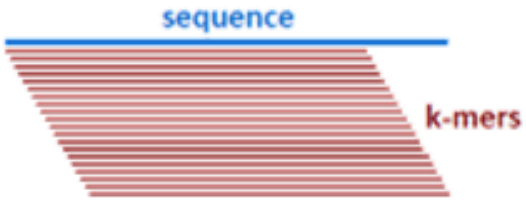
UniUDLab and Administration
GiusiZaina
Nicoletta Felice
Cristian DelFabbro

Supported by the ERC project NOVABREED -Novel variation in plant breeding and the plant pan-genomes (Grant agreement no.: 294780)

K-mer frequency based error correction algorithm



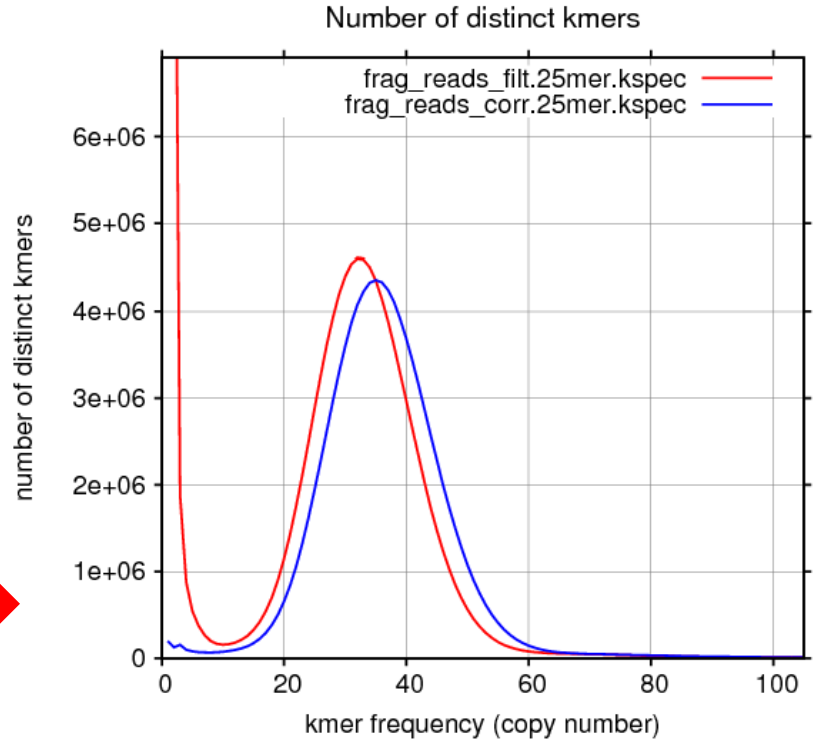
calculate k-mer frequency spectrum



apply corrections to reads

change the sequence to the closest

```
CCGGATTTACGCCGAGTGA x 43
CCGGATTTACGCTGAGTGA x 1
                   ^
```



Assembly metrics

1.Order contigs/scaffolds in descreasing length

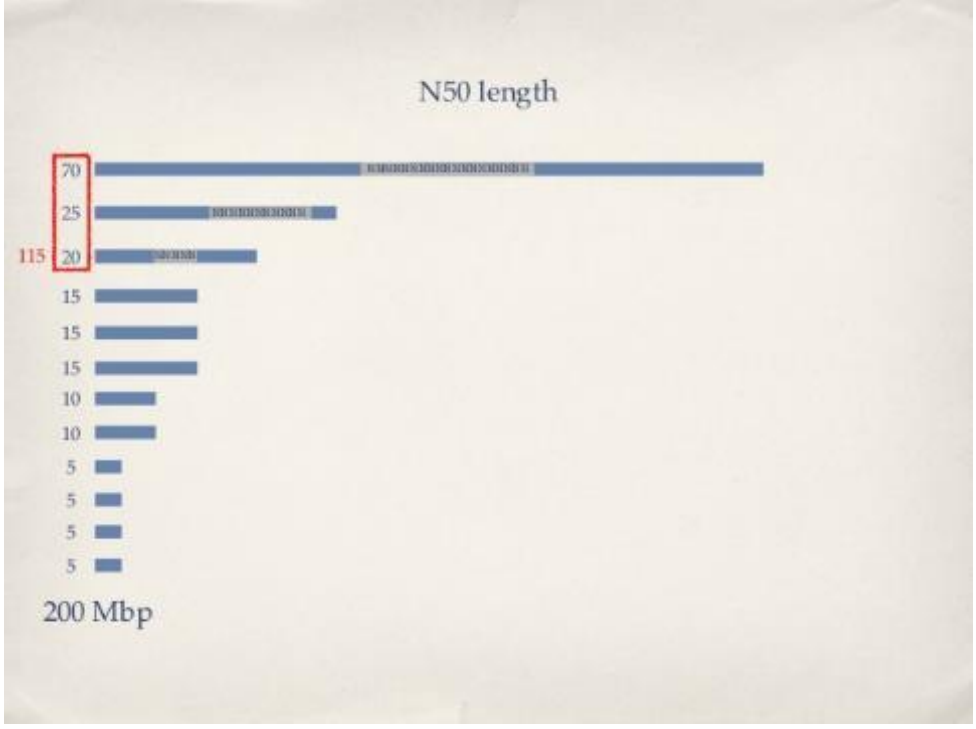
2.Calculate contigs total length *(TOT_LEN)*

N50 = first *n* given $\sum_{i=1}^{n} length(i) \geq TOT_{LEN}/2$

L50 = length of the last contig/scaffold added to cover 50% or more of the assembly

NG50 = first *n* then cui
$\sum_{i=1}^{n} length(i) \geq GENOME_{LEN}/2$

LG50 = …

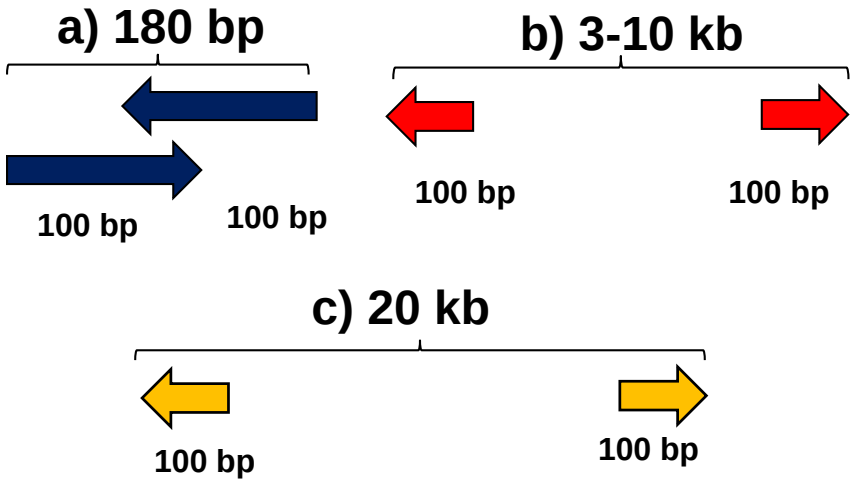


The ALLPATHS-LG receipt

a) ~45X Overlapping fragment libraries

b) ~45X Jumping libraries

c) ~1X Long Jumping libraries



a) ~30X Overlapping fragment libraries

b) ~20X Non-overlapping fragment libraries

c) ~20X 3kb Jumping libraries + ~10X 10kb (Long?) Jumping libraries