Effects Of Filtering, Trimming, Sampling And K-mer Value On *De Novo* Assembly Of Illumina GA Reads.

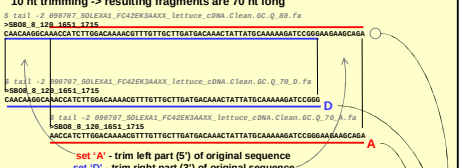
Alexander Kozik , **Marta Matvienko** , **Richard W. Michelmore**. The Genome Center, University of California, Davis, CA, 95616, USA.

NSF Plant Genome Program award #0820451

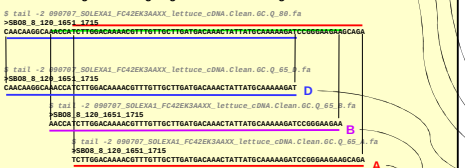
Assembling millions of short reads generated by the Illumina Genome Analyzer (IGA) sequencing platform is a challenge. Multiple algorithms have been and are being developed for *de novo* assembly of DNA fragments generated by the IGA. We are developing protocols and optimizing parameters for assembling the transcriptome and gene space of the three gigabase genome of lettuce. The selection of high quality reads is a key step for successful assembly. We have analyzed the consequences of filtering, trimming, and sampling for the Velvet assembler. K-mer value is being optimized using subsets of variable read lengths. Velvet and CLC (CLC DNA Workbench) assemblers are being compared using the same input datasets. Contigs produced by CLC and Velvet were assembled into a "super-assembly" using the CAP3 program and visualized with our custom Python Contig Viewer. Assemblies were validated by comparison with known conserved sequences of other species and by alignment of the transcriptome assembly to the gene space assembly. This analysis is part of the Compositae Genome Project http://compgenomics.ucdavis.edu/. Our protocols and tools to process and filter Illumina GA reads for *de novo* assembly are available at http://code.google.com/p/atgc-illumina/.

Trimming and Sampling of mRNA-Seq IGA reads

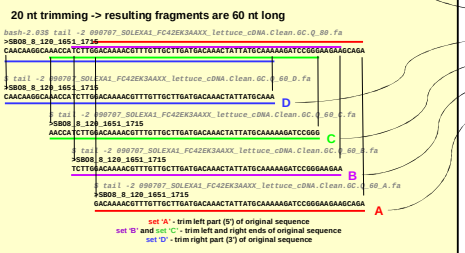
10 nt trimming -> resulting fragments are 70 nt long

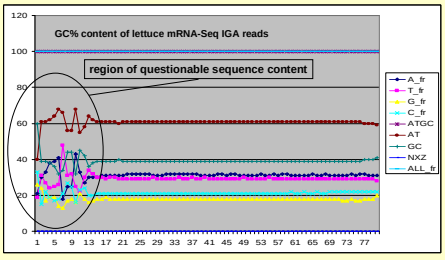


15 nt trimming -> resulting fragments are 65 nt long



20 nt trimming -> resulting fragments are 60 nt long

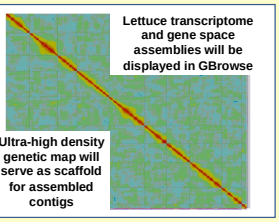




Project Objective:

Optimization of assembly of short reads from Illumina Genome Analyzer, with an emphasis on mRNA-Seq data from lettuce transcriptome

1. Effect of trimming on Velvet assembly
2. Effect of K-mer values on Velvet assembly
3. Large-scale assembly with Velvet
4. CAP3 super-assembly of Velvet assemblies



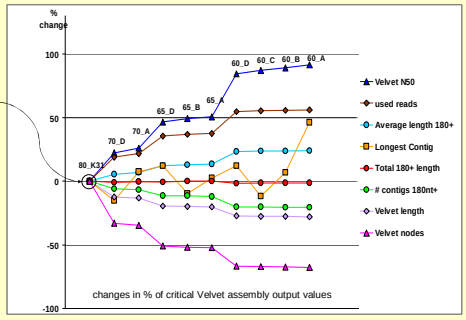
Lettuce transcriptome and gene space assemblies will be displayed in GBrowse

Ultra-high density genetic map will serve as scaffold for assembled contigs

Effect of trimming on transcriptome assemblies

sample	# of reads (million)	used reads (million)	velvet N50	velvet # nodes	velvet length million nt	longest	median 180 nt +	average 180 nt +	contigs 180nt +	total 180+ length million nt
80_K27	22.62	8.91	126	968,002	53.08	2428	260	312	79,889	24.94
80_K29	22.62	9.78	141	829,641	50.38	2448	262	318	82,619	26.28
80	**80_K31**	**22.62**	**10.73**	**156**	**719,028**	**48.97**	**4417**	**267**	**326**	**85,123**	**27.76**
70_D	70_D	22.62	12.78	191	483,450	43.01	3752	278	344	80,045	27.53
70_A	**70_A**	22.62	**13.08**	**197**	469,949	42.62	**4752**	279	**349**	**79,578**	**27.74**
65_D	65_D	22.62	14.53	229	353,683	39.51	4964	290	366	75,602	27.64
65_B	65_B	22.62	14.68	233	347,461	39.28	3995	289	368	75,387	27.76
65_A	**65_A**	22.62	**14.75**	**235**	343,477	39.11	**4523**	290	**370**	**75,038**	**27.76**
60_D	60_D	22.62	16.61	288	239,076	35.71	4964	307	402	67,942	27.29
60_C	60_C	22.62	16.67	292	237,207	35.57	3900	305	403	68,112	27.41
60_B	60_B	22.62	16.71	295	235,140	35.44	4720	306	404	67,836	27.41
60_A	**60_A**	22.62	**16.76**	**299**	233,365	35.33	**6459**	306	**405**	**67,744**	**27.43**

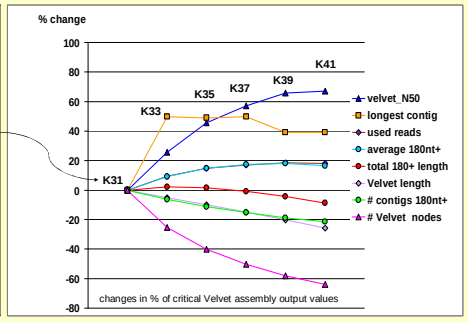
Assembly of 22.62 million mRNA-Seq reads of variable length with fixed 31 K-mer value



Impact of Velvet K-mer on transcriptome assemblies

sample	# of reads (million)	used reads (million)	velvet N50	velvet # nodes	velvet length million nt	longest	median 180 nt +	average 180 nt +	contigs 180nt +	total 180+ length million nt
80_QC K31	47.50	17.03	126	1,160,217	60.85	4272	261	318	93,175	29.67
60_QC K31	47.50	33.44	303	325,378	42.83	5131	313	422	76,994	32.55
60_QC K33	47.50	36.64	381	242,908	40.68	7682	328	461	72,092	33.26
60_QC K35	47.50	38.39	441	194,847	38.58	7640	338	484	68,241	33.03
60_QC K37	47.50	39.17	476	162,163	36.46	7682	342	495	65,230	32.26
60_QC K39	**47.50**	**39.58**	**502**	**136,282**	**34.20**	**7141**	**343**	**498**	**62,557**	**31.16**
60_QC K41	47.50	39.47	507	117,446	31.76	7141	338	491	60,446	29.68

Assembly of 47.50 million mRNA-Seq reads of fixed 60 nt length with variable K-mer value



Large-scale whole transcriptome assembly: Assembly of 77.5 million reads with Velvet and downstream assembly of these assemblies with CAP3 into 'super-assembly'

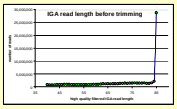
Initial Dataset – 77.5 million mRNA-Seq reads

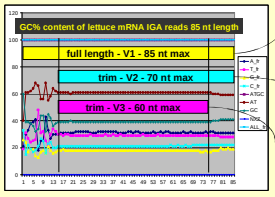
subsets with trimming options:

V1 - reads within range 40 - 85 (full filtered length 'as is')

V2 - reads within range 25 - 70 (trim first 15 nt)

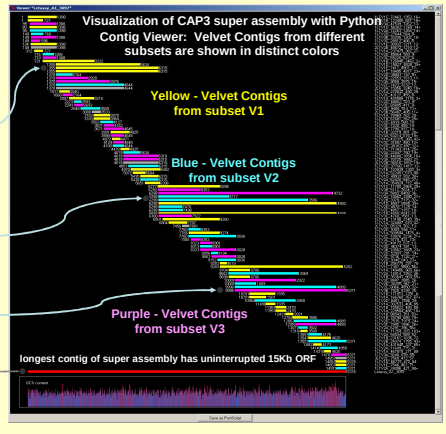
V3 - reads within range 25 - 60 (trim first 15 and last 10 nt)





Several rounds of Velvet assembly with different parameters and sampling

sample	# of reads (million)	used reads (million)	velvet N50	velvet # nodes	velvet length million nt	longest	median 300 nt +	average 300 nt +	contigs 300 nt +	total 300+ length million nt
V1_K35	77.5	15.8	138	1,097,943	63.8	3745	413	489	41,321	20.2
V1_K41(*)	77.5	31.5	251	514,297	54.3	6673	453	562	51,068	28.7
V1_K47	77.5	47.5	482	230,756	47.1	8711	522	691	50,186	34.7
V1_K51	77.5	49.6	580	166,477	43.3	10258	544	729	47,486	34.6
V1_K55	60.9 [*]	46.8	653	124,677	39.6	10843	557	749	44,852	33.6
V2_K37	77.5	39.4	367	290,027	45.4	6547	496	634	45,608	28.9
V2_K41(*)	77.5	45.0	496	186,892	41.2	10139	532	697	43,208	30.1
V2_K45	77.5	45.5	568	133,766	37.1	7170	547	719	40,457	29.1
V2_K49	60.9 [*]	43.4	589	100,648	32.6	7141	542	709	37,677	26.7
V3_K33	60.9 [*]	35.8	361	292,805	44.1	8441	499	640	42,805	27.4
V3_K39	60.9 [*]	44.8	536	154,636	37.6	10394	548	717	39,041	28.0
V3_K43	60.9 [*]	46.0	565	112,506	32.8	7141	546	712	35,948	25.6



CAP3 assembly of different subsets of Velvet contigs - superassembly

Input for super assembly:

519,657 contigs 300 nt or longer after 12 iterations of Velvet assembly

Total number of non-redundant Velvet contigs in 12 subsets - 428,122

CAP3 assembly with non-redundant set (parameters: 80nt overlap, 95% identity)

Super assembly output - 58,142 unigenes:
40,853 new CAP3 contigs and 17,289 non-assembled Velvet contigs

longest contig: 15,376 nt - uninterrupted 15 Kb ORF
total length of super assembly: 52,397,565 nt
average contig length: 901 nt
median: 685 nt

BLASTX analysis - super assembly versus Plant Reference Database
1e-10 expect cutoff: 42,633 contigs (73%) are with hits to known proteins

16,446 long contigs (1,000nt+) out of 17,000 (97%) have hits to known protein sequences

Validation of superassembly (Velvet + CAP3) by BLAST search versus Plant Reference Database

Query ID (super assembly Contig ID)	Subject (BLAST hit) ID	BLAST hit description (subject annotation)	percent identity	alignment length	# gaps	query start (nt)	query end (nt)	subject start (nt)	subject end (nt)	query length (nt)	subject length (nt)
Letassy_A1_3097	Arab_thal.NP_186875	[NP_186875.2 GI:30678519] { BIG (BIG); binding / ubiquitin-protein ligase/ zinc ion binding } [Arabidopsis thaliana]	63	3255	5159	3	204	15194	1	5096	15376	5098
Letassy_A1_290	Vitl_vinI.XP_002274451	[XP_002274451.1 GI:225432136] { PREDICTED: similar to vacuolar protein sorting 13C protein-like } [Vitis vinifera]	63	2731	4324	3	12718	65	42	4285	12879	4286
Letassy_A1_7087	Arab_thal.NP_197702	[NP_197702.1 GI:15237223] { zinc finger (C3HC4-type RING finger) family protein } [Arabidopsis thaliana]	57	2448	4264	3	1	12573	356	4597	12573	4706
Letassy_A1_11111	Vitl_vinI.XP_002268896	[XP_002268896.1 GI:225461945] { PREDICTED: similar to E3 ubiquitin-protein ligase UPL1 } [Vitis vinifera]	69	2683	3855	-3	11129	45	1	3754	11478	3754
Letassy_A1_38867	Vitl_vinI.XP_002283711	[XP_002283711.1 GI:225459044] { PREDICTED: hypothetical protein } [Vitis vinifera]	73	2727	3730	3	284	11092	3	3653	11127	3653
Letassy_A1_94	Popu_tric.XP_002327756	predicted protein [Populus trichocarpa]	81	2826	3466	3	1	10248	459	3881	10451	3881
Letassy_A1_37088	Vitl_vinI.XP_002274761	[XP_002274761.1 GI:225437162] { PREDICTED: hypothetical protein } [Vitis vinifera]	71	2414	3390	-3	9949	2	1	3325	10362	4546
Letassy_A1_4189	Vitl_vinI.XP_002283711	[XP_002283711.1 GI:225459044] { PREDICTED: hypothetical protein } [Vitis vinifera]	65	2443	3734	-3	10215	151	3	3653	10262	3653

Conclusions and Implications:

For mRNA-Seq, it is important to trim the initial ~15 bases. (This may not be important for genomic reads; data not shown.)

Filtered high-quality 60 nt mRNA-Seq fragments assemble better using Velvet than raw 85 nt reads.

Current results indicate that optimal Velvet K-mer value is ~2/3 of the fragment length within the range of 60 to 80 nt (subject of further investigation).

Velvet assemblies of different subsets with variable options may contribute different assemblies to generate full-length super assemblies regardless of optimal K-mer values and the trimming of reads.

Comparison of Velvet with different extended run parameters and the commercial CLC assembler on mRNA-Seq datasets is the subject of on-going investigation.

Additional Information:
Further information about assembly procedure and side-by-side comparison of Velvet with CLC assembler is available online at http://code.google.com/p/atgc-illumina

For lettuce gene space sequencing and assembly, see poster P014 Matvienko et al.: *Integration Of Genetics And Sequencing Of The Transcriptome And Gene Space Of Lettuce*