Metagenome extraction pipeline

Lucas Sinclair, lucas.sinclair@me.com

Last modified:

Mon Nov 04 2013

All numbers are fake
This document
describes a plan, it
is not scientific work.

0) Start: with N metagenome samples that are related in some way (different extraction method, different times, different locations). In this example: different locations.

Metagenomes (samples)
Halsjärvi (AE09)
Mekkojärvi (AE12)
Alinen Mustajärvi (AE13)

1) Clean: the data in each sample with specific thresholds on PHRED quality scores. Maybe sliding window strategy ?

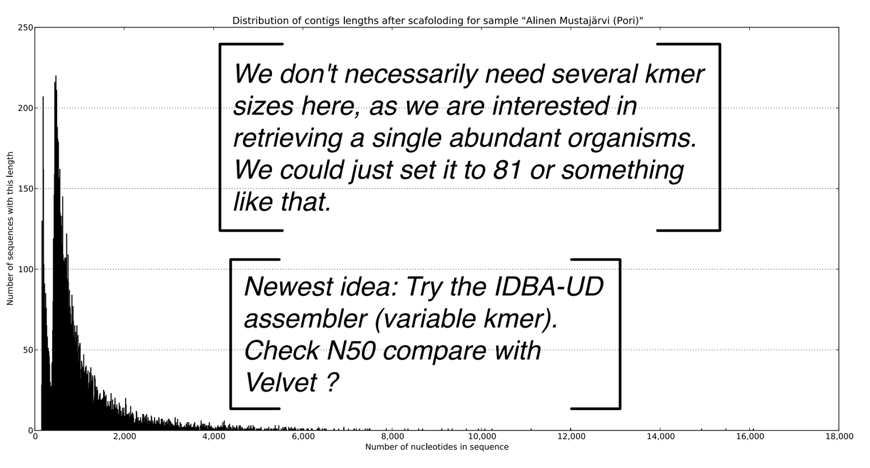
#	Pairs before	Pairs after	Loss
AE09	89'746'032	87'746'010	3%
AE12	43'926'191	40'926'456	5%
AE13	46'119'110	42'113'230	17%

Sickle

2) Co-assemble: them all together on the Halvan machine or on a Cray using Ray (accessible via SLURM). The result is a bunch of contigs.

Should we use the khmer program to even out the coverage since we are using MDA in the lab?
Maximum for a 1000 Cray: approx. 400'000'000 pairs

Contigs
Contig C1) length: 10'390
Contig C2) length: 9'780
Contig C3) length: 8'400
Contig C4) length: 8'125
Contig C5) length: 7'903
Contig C6) length: 7'842
Contig C7) length: 7'491
Contig C8) length: 6'233
Contig C9) length: 6'127
ETC...



We don't necessarily need several kmer sizes here, as we are interested in retrieving a single abundant organisms. We could just set it to 81 or something like that.

Newest idea: Try the IDBA-UD assembler (variable kmer). Check N50 compare with Velvet ?

Ray

3) Coverage: For every contig, we are going to compute the coverage it had in every one of our different samples.

#	length	AE09	AE12	AE13
C1	10'567	1110	501	2278
C2	9'780	543	1578	5031
C3	8'400	1000	478	2104
C4	8'125	2	6	7
C5	7'903	2140	1023	4456
C6	7'842	10567	20345	40842
C7	7'491	100	54	234
C8	6'233	34	56	23
C9	6'127	12	34	66

We do this by mapping all the reads of each metagenome successively to the set of contigs we have from the co-assembly. We can use the Bowtie algorithm for doing this or maybe Bowtie ? How do we deal with reads that map to several places ? Count zero or count one for each ?

Bowtie
MarkDuplicates
Samtools
BEDtools
Custom parsing

4) Normalize: this information by the length of the config and convert the number into coverage.

#	length	AE09	AE12	AE13
C1	10'567	30x	15x	60x
C2	9'780	7x	9x	6x
C3	8'400	31x	14x	62x
C4	8'125	17x	45x	67x
C5	7'903	29x	16x	62x
C6	7'842	120x	223x	600x
C7	7'491	4x	2x	4x
C8	6'233	5x	5x	3x
C9	6'127	1x	2x	3x

Points that are close in this N dimensional space are likely to belong to the same genomic unit (species and/or strain).

This means coverage is similar across species (columns) not across contigs (rows).

n/a (in pipeline)

5) Tetra nucleotide frequencies: are computed for every contig. This results in a vector that has 256 dimensionality.

#	length	AE09	AE12	AE13	Sequence freq.
C1	10'567	30x	15x	60x	<1.4;6.4;3.5;...>
C2	9'780	7x	9x	6x	<2.1;5.2;7.1;...>
C3	8'400	31x	14x	62x	<0.1;9.2;7.6;...>
C4	8'125	17x	45x	67x	<1.4;6.4;3.5;...>
C5	7'903	29x	16x	62x	<9.4;4.4;1.5;...>
C6	7'842	120x	223x	600x	<2.1;5.2;7.1;...>
C7	7'491	4x	2x	4x	<0.1;9.2;7.6;...>
C8	6'233	5x	5x	3x	<2.1;5.2;7.1;...>
C9	6'127	1x	2x	3x	<4.1;9.2;7.6;...>

Idea: Compare with a binning method that only uses sequence frequencies and not the coverage information such as the self-organizing maps?

n/a (in pipeline)

6) Pairdness: We should also exploit the pairdness information at some point. Maybe by taking every contig successively, and going through its reads counting how many mates fall in an other contig ?

This will create a matrix where, of course the diagonal should be filled with very high numbers.

Ideally, we only need to look at reads mapping close to both ends of the contigs to determine which one should be the neighbor to the 5' side and the 3' side.

How should we use this extra info ? We can make a list of most probable linked contigs and add it to the rest of the information? Or maybe we should only use this information at the end for checking/ manually curating ?

#	C1	C2	C3	C4	C5	C6	...
C1	1'000'567	1	89	0	112	1	...
C2	1	2'909'567	1	1	1	0	...
C3	45	1	6'000'567	1	44	1	...
C4	2	1	1	3'456'567	1	1	...
C5	87	0	12	1	2'010'567	1	...
C6	1	1	1	1	0	1'000'567	...
C7	1	1	1	1	1	1	...
C8	1	0	1	0	1	0	...
C9	1	1	0	0	1	1	...

n/a (in pipeline)

#	length	AE09	AE12	AE13	Sequence freq.	Best pairs
C1	10'567	30x	15x	60x	<1.4;6.4;3.5;...>	C5R, C3L, C2R, ..
C2	9'780	7x	9x	6x	<2.1;5.2;7.1;...>	C9L, C11R, C122L, ..
C3	8'400	31x	14x	62x	<1.3;5.9;5.6;...>	C1L, C5R, C94L, C111L..
C4	8'125	17x	45x	67x	<1.4;6.4;3.5;...>	C17R, C56L, C85R, ...
C5	7'903	29x	16x	62x	<1.9;5.2;1.5;...>	C1R, C3R, C91L, C99R..
C6	7'842	120x	223x	600x	<2.1;5.2;7.1;...>	C56L, C12L, C43R, ...
C7	7'491	4x	2x	4x	<0.1;9.2;7.6;...>	C41R, C11R, C981R, ...
C8	6'233	5x	5x	3x	<2.1;5.2;7.1;...>	C51R, C22L, C33R, ...
C9	6'127	1x	2x	3x	<4.1;9.2;7.6;...>	C56R, C12L, C43L, ...

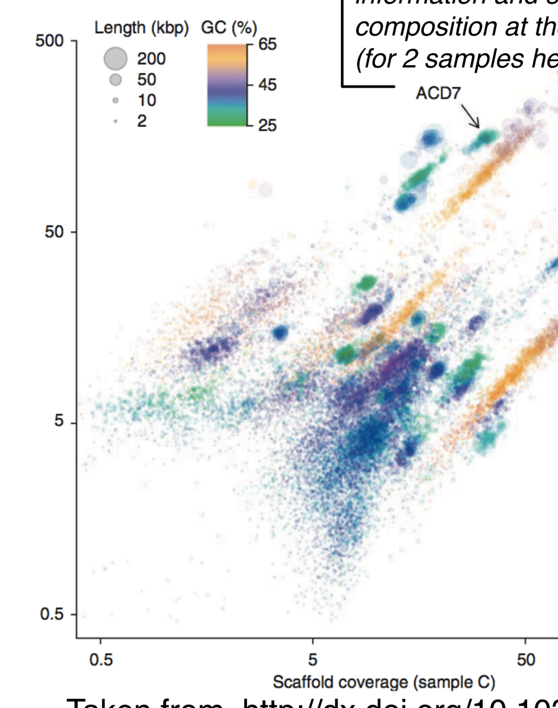
Keep the pipeline open for adding extra information to this matrix such as distribution of environment variables extracted from the recruitment of reads to each of the contigs.

7) Clustering: Now we are ready to form clusters. This is a point that still needs to be determined. Choose some algorithm from http://scikit-learn.org ? Maybe some hierarchical clustering ?

The result is a list of bins each containing a number of contigs.

Bins
B1 = C1+C3+C5
B2 = C2+C4
B3 = C6+C7+C8+C9
ETC...

Here is a way to view coverage information and sequence composition at the same time (for 2 samples here):



Taken from http://dx.doi.org/10.1038/nbt.2579

Hierarchical clustering package in R

8) Assignment: once we have a cluster of contigs (perhaps one that cumulatively represents 20% in the first sample and 25% in the second) how do we know what species it is ? Will there be a 16S present ? What marker genes can we use ? How do we resolve their position on a phylogenetic tree ?

Bins with associated taxonomy
B1 = C1+C3+C5 (Candidate div. OD1)
B2 = C2+C4 (Polynucleobacter)
B3 = C6+C7+C8+C9 (Pelodictyon)
ETC...

Just blast all the contigs against NR ?

Lorem ipsum

9) Reassembly: Once we have decided which reads from the original sets of all reads go together, we reassemble them (possibly with lower kmer number) producing the same bins (now=genomes) but possibly new number of contigs (now=scaffolds)

#	Scaffolds	Taxa
G1	S1, S2	Candidate div. OD1
G2	S3, S4, S5	Polynucleobacter
G3	S6, S7	Pelodictyon
.	...	...

10) Validation: we need a way to estimate the completeness of the genome produced. What strategy to go for ?

#	Scaffolds	Taxa	Completeness
G1	S1, S2	Candidate div. OD1	95%
G2	S3, S4, S5	Polynucleobacter	82%
G3	S6, S7	Pelodictyon	73%
.	...	...	...

Could we count the number of tRNAs ? Number of 35 single copy house keeping genes ?

For instance could we check the number of ORF frame shifts, this could give us a measure of how many scaffolds are chimeric ?

Lorem ipsum

11) Comparison: We have other related OD1s:
- Banfield publication
- Our epilimnion SAG
- SAGs from other publication

How can we compare the one we now built with them in a meaningful way ? MUMmer ?

#	Scaffolds	Taxa	Completeness	OD1-A	OD1-B	OD1-C	...
G1	S1, S2	Candidate div. OD1	95%	99%	88%	92%	...
G2	S3, S4, S5	Polynucleobacter	82%	12%	15%	1%	...
G3	S6, S7	Pelodictyon	73%	13%	5%	23%	...
.	...	...	...	...	...	...	...

We could also combine step 10 and 11 with a comparison depedamt validation: Determining what is common amongst the already published SAGs, try to find it in our own contig cluster and evaluate completeness ?

12) Annotation: We need to annotate the genome created.

For instance, we can plug in the results to an ORF predictor, blast against COG, the Metacyc database, and view the results in pathways tools ?

What other automated annotation tools are there ?

#	length	Number of ORFs
S1	15'012	240
S2	12'945	125
.	...	...

We want to identify specific pathways, and which organisms have them or not.

13) Metabolic potential: We should evaluate its metabolic potential and make a cell digram representing what it can do and cannot do. This is the ultimate goal of the whole procedure.

#	length	AE09	AE12	AE13	Sequence freq.	Best pairs	Predicted Taxonomy?	16S found or not HMM
C1	10'567	30x	15x	60x	<1.4;6.4;3.5;...>	C5R, C3L, C2R, ..	OD1	False
C2	9'780	7x	9x	6x	<2.1;5.2;7.1;...>	C9, C11, C122, ..	Bacteriodites	False
C3	8'400	31x	14x	62x	<1.3;5.9;5.6;...>	C1, C5, C94, C111..	OD1	True: OD1
C4	8'125	17x	45x	67x	<1.4;6.4;3.5;...>	C17, C56, C85, ...	OP11	False
C5	7'903	29x	16x	62x	<1.9;5.2;1.5;...>	C1, C3, C91, C99..	Chlorobi	False
C6	7'842	120x	223x	600x	<2.1;5.2;7.1;...>	C56, C12, C43, ...	Blah	False
C7	7'491	4x	2x	4x	<0.1;9.2;7.6;...>	C41, C11, C981, ...	Blah	False
C8	6'233	5x	5x	3x	<2.1;5.2;7.1;...>	C51, C22, C33, ...	Blah	True: OP11
C9	6'127	1x	2x	3x	<4.1;9.2;7.6;...>	C56, C12, C43, ...	Blah	False