

Iteration	Purpose	Defined	Due	Delivered	Version	Notes
1	Initial with basic infrastructure	19/02/2010	12/03/2010	12/03/2010	1.00	
2	File I/O and data structures	12/03/2010	01/04/2010	06/04/2010	1.01	
3	AYB Initialisation	08/04/2010	26/04/2010	27/04/2010	1.02	
4	Base Call Loop	06/05/2010	21/05/2010	20/05/2010	1.03	
5	MPN, quality output, datablock	28/05/2010	17/06/2010	17/06/2010	1.04	
6	P Solver, MPN unit test, cif format Pause for testing and experimentation	22/06/2010	29/07/2010	30/07/2010	1.05	
7	Final working values, mu arg, bad data handling	25/08/2010	21/09/2010	22/09/2010	1.06	

Index	Priority	Task	Planned	Actual	Notes
1		Initial system: Make; Run	1	1	
2		Program Version	1	1	
3		Program arguments: Read and store: Infrastructure	1	1	
4		Program Log: Infrastructure	1	1	
5		Signal Handler	1	1	
6		File I/O: Locate, open/create, read/write, close	2	2	Requires args input (i), output (o);
7		Data Structures: Tile, Cluster, Matrix	2	2	
8		Intensities Input: Locate, read, store, tidy up	2	2	
9		Matrix Input: Locate, read, store, tidy up; Read M, N, P	2	3	Requires args M (M), N (N), P (P)
10		Processed Intensities: Calculate and store, tidy up	3	3	
11		Initial Sequence: By maximum intensity	3	3	
12		Sequence output: Create file, write data, close	3	3	
13		Replace message macros with vfprintf	2	2	
14		Expand message severity list to include debug	2	2	
15		Configure Doxygen; add comments to new HM files	2	2	
16		Matrix: Add read_methods for multiple styles	3	3	1) Intensities 2) As rows of columns. Move matrix read from cluster.
17		Initial Lambda: Ignore weights	3	3	

18	Calculate Covariance	4	4	
19	Estimate Lambda	4	4	
20	Call Bases	4	4	
21	Base Call Loop	4	4	Requires args niter (n?); change ncycles to c?
22	Utility: Reimplement xfree null return	4	4	Use ** or return, could apply to all free_* functions
23	XIO: Check and amend file structure handling; nulls etc.; null return on close	4	4	Make interface as close to normal file handle as possible
24	XIO: rename initialise_aybstd to not contain 'ayb'	4	4	
25	I/O: If input file open fails then warn and carry on	4	4	But stop if output fails as all will
26	I/O: Output files should not be compressed	4	4	Test: compressed input, input with no extension, input with no delimiter, combo?
27	I/O: Create output directories if do not exist	4	4	Include log; abort if create fails.
28	I/O: Intensities: Expand fixed match to “_int.txt”	4	4	As original
29	Program log: Replace fixed “ayb” with prefix	4	4	hhmm not sufficient for parallel runs
30	Program log: Add initial information line	4	4	Program name, date/time
31	Program log: switch order of warning and information	4	4	

32	Process Intensities file for analysis by block	5	5	New argument blockstring (b) of the form: RnInCn, decoded as: R=>read I=>ignore C=>concatenate to previous block Note no difference in analysis for forward and backward data. Then: less data than specified=>abort program; more data than specified=>warn and continue
33	Implement Quality Scores	5	5	Actually 2 alternative outputs fasta/fastq. Requires arg format (f).
34	Implement MPN estimation.	5	5	Get from AYBc; convert intensities access from single array to via cluster list. Use MPN initialisation as in AYBc
35	Allow selection of routines to solve for P. Remove simultaneous solve for P and N (from TM)	6	6	Alternatives are: Standard SVD; Standard SVD then zero negative entries; Non-negative least squares.
36	Unit test of MPN estimation.	6	6	Already exists in an old form but requires changes; coercing values into array/cluster/tile; other bits.
37	Option to read Intensity data in cif format.	6	6	Get from AYBc; convert intensities access from single array to via cluster list. Requires arg dataformat (d).
38	Optional output of final processed intensities, M, P, N, lambda, weights	7	7	Final processed format as intensities input, illumina or cif. Write cif format already exists. 1 file per input. Requires arg working (w)
39	mu should be a parameter as in AYBc.	7	7	Requires arg mu (m)
40	Introduce a diagonal delta to the solver routines (ridge regression); avoids failure to solve when data bad .	7	7	Use value 1 which is small compared with typical matrix values.
41	Count zero lambdas for each file and output at end if any.	7	7	Output per iteration.
42	Stop and issue message if data error detected.	7	7	Initially failure to calculate covariance.
43	Remove padding zeros from output cluster line.	7	7	There are typically > 10,000 clusters but could be any number.
44	Couple of bugs in string length allocation.	7	7	Cause segment errors.

45	Make double calculations standard.	7	7	Accuracy of float not sufficient for larger values.
----	------------------------------------	---	---	---

## New

Method of filtering the list of clusters according to some criteria; resultant finds should be marked (separately)

Possible filters:

- cycle in cluster with all signals zero (signals missing)
- cycle in cluster with signal greater than some threshold (argument)

The filter\_list, filtercopy\_list and split\_list functions do this but may need to be modified to retain the same ordering of elements (cf read\_TILE and read\_known\_TILE).

Look at efficiency.

Automate module testing with a test script. Create reference results for future comparison.

Enhancements to cluster/tile; null values, cif.

## Errors

In Matrix, test instructions are wrong way round

Functions that take no arguments should include void inside the argument brackets.

## Consider

Check all algorithms use preferred versions.

e.g. calculate lambda.

AYB struct: Make bases and quals part of cluster, also lambda and weights?

But will make tile bigger with empty space? weights are accessed as an array for mean/var purpose; also use of set/scale\_MAT.

Parameters to makefile to allow debug/double?

## Refactor

- Split ayb\_model
- Rationalise use of int/long/uint32 types
- Optionally use message in matrix/cluster/tile/nuc/more? controlled by compiler switch.
- New message functionality to return message string
- Change message to use CSTRING instead of fixed length
- Message: move message file check to filename routine
- Message: use a different word than type for msgtype?
- Add segmentation violation handler
- ayb\_options: move structure to source file, add get for each item
- Add descriptions to func declarations in headers
- Make --version option comply to GNU coding standards

Possibly extract all AYB into new ayb. It's possible that static variables are inefficient - may want to pass AYB pointer in but keep internals hidden.  
Be aware of efficiency issues.

Set switch flag in message?

For use with perror, length of string? vsnprintf?

So can return default

Confusing in documentation

Only when there is an item to get, or remove altogether.

Not needed with Doxygen?

Consider

ayb\_model: make bases & quals a list (per cluster)

ayb\_model: does AYB struct need to be public at all? or even the standard functions?

\* xio, matrix, cluster, tile not yet documented

\* tidy dirio.h on next change ???

Index	Note	Resolution
1	Divide by zero does not cause FPE so not tested	
2	Have not resolved all the get memory issues during data append.	



Index	Task	Additional Requirements	Completed	Notes
38	Optional output of final processed intensities, M, P, N, lambda, weights	Final processed format as intensities input, illumina or cif. Write cif format already exists. One file per input with tag "pif". Single file (tag "final") used for all other values in show_MAT format. Requires arg working (w)	20/09/2010	Matrix write_to_line; Cluster write_coords; new read_first to get lane & tile, common read parts to subproc; Tile write_lane_tile, use cluster read_first. Cif set_from_real, create_cif, option to show all; Utility new real_t round function definition. Ayb_model standard (write each line) or cif (store and write at end) output; open/write/close_processed; pass last iter flag to estimate_bases; option to show only part AYB structure; new input format info message.
39	mu should be a parameter as in AYBc.	Requires arg mu (m)	21/09/2010	Adjusts range of quality scores; smaller value produces higher maximum quality score.
40	Introduce a diagonal delta to the solver routines (ridge regression); avoids failure to solve when data bad .	Use value 1 which is small compared with typical matrix values.	26/08/2010	Value stored in ayb_model as constant DELTA_DIAG.
41	Count zero lambdas for each file and output at end if any.	Output per iteration.	25/08/2010	Store in ayb_model. Create string for message with count for each iteration.
42	Stop and issue message if data error detected.	Initially failure to calculate covariance.	26/08/2010	New call_base_null, same as used for zero lambda. estimate_bases can return err val.
43	Remove padding zeros from output cluster line.	There are typically > 10,000 clusters but could be any number.	21/09/2010	In ayb_model output_results.
44	Couple of bugs in string length allocation.	Cause segment errors.	25/08/2010	In dirio: output_name(_cif), calculation of newname length requires bracket around blk query part or result seems to always be 1; scan_inputs fixed buffer insufficient.
45	Make double calculations standard.	Accuracy of float not sufficient for larger values.	25/08/2010	Rename original makefile to makefloat (not controlled).

Index	Task	Additional Requirements	Completed	Notes
35	Allow selection of routines to solve for P. Remove simultaneous solve for P and N (from TM).	Alternatives are: Standard SVD; Standard SVD then zero negative entries; Non-negative least squares. New arg solver (S).	09/07/2010 28/07/2010	Changes to help, model, mpn, matrix error. Need fortran compiler switch out (NFORTTRAN) for eclipse.
36	Unit test of MPN estimation.	Already exists in an old form but requires changes; coercing values into array/cluster/tile; other bits.	23/07/2010	coerce matrix and expanded to array/cluster/tile. Calculate intermediates and new values.
37	Option to read Intensity data in cif format.	Get from AYBc; convert intensities access from single array to via cluster list. Requires arg dataformat (d).	21/07/2010	Adopt cif. Store input format in dirio; alternatives for file search and output filename. New read_cif_tile/cluster. In model alternative read and output filename. New debug output constraints.

Index	Task	Additional Requirements	Completed	Notes
32	Process Intensities file for analysis by block	New argument blockstring (b) of the form: InRnCn, decoded as: I=>ignore R=>read C=>concatenate to previous block Note no difference in analysis for forward and backward data. Then: less data than specified=>abort program; more data than specified=>warn and continue Add letter extension to output file name if more than one block.	09/06/2010	Replace option cycles (c) with blockstring (b). New datablock class for datablock structure. Create tile/cluster/matrix append functions. Pre-process input tile and create an array of sub-tile pointers. Add ncycle to TILE structure. No longer need ncycle return from tile. Change action if not enough cycles; message if spare cycles (pass flag on to read matrix line to check if first line). In read_tile error if later clusters have < cycles than previous; indicates a faulty file. New open_output_blk adds a block suffix letter. Analysis now a loop.
33	Implement Quality Scores	Actually 2 alternative outputs fasta/fastq. Requires arg format (f).	16/06/2010	Function match_string to utility.
34	Implement MPN estimation.	Get from AYBc; convert intensities access from single array to via cluster list. Use MPN initialisation as in AYBc.	15/06/2010	Need new modules mpn, statistics; new functions update_cluster_weights, estimate_MPN (model), expected_intensities (intensities), some matrix and lapack. Read-in matrices become optional; new dirio func to say matrix specified.

Index	Task	Additional Requirements	Completed	Notes
18	Calculate Covariance		11/05/2010	Put in ayb_model not call_bases because uses structure of AYB. Scale reciprocal removed.
19	Estimate Lambda		12/05/2010	Use estimate_lambdaWLS as originally described.
20	Call Bases		11/05/2010	Constant Mu used for quality score.
21	Base Call Loop	Requires args niter (n); change ncycles to c	12/05/2010	Niter static in ayb_model.
22	Utility: Reimplement xfree null return	Use ** or return, could apply to all free_* functions	13/05/2010	Used by free_AYB/CSTRING/(MAT)/CLUSTER/TILE. Leave xfree returning void and instead return null pointer from free_functions. Do not have to have a return in call to free_x if not needed, e.g. a local var. ARRAY/LIST never freed where return matters.
23	XIO: Check and amend file structure handling; nulls etc.; null return on close	Make interface as close to normal file handle as possible	17/05/2010	Free structure and return null pointer if open fails. Return null pointer on close (different from normal file handle operation)
24	XIO: rename initialise_aybstd to not contain 'ayb'		13/05/2010	
25	I/O: If input file open fails then warn and carry on	But stop if output fails as all will	17/05/2010	Dirio open_output: loop until successful open.
26	I/O: Output files should not be compressed	Test: compressed input, input with no extension, input with no delimiter, combo?	18/05/2010	
27	I/O: Create output directories if do not exist	Include log; abort if create fails.	18/05/2010	Exist, no exist, exist file not dir.
28	I/O: Intensities: Expand fixed match to "_int.txt"	As original	17/05/2010	
29	Program log: Replace fixed "ayb" with prefix	hhmm not sufficient for parallel runs	19/05/2010	
30	Program log: Add initial information line	Program name, date/time	19/05/2010	AYB Message Log; user name and datetime
31	Program log: switch order of warning and information		19/05/2010	

Index	Task	Additional Requirements	Completed	Notes
9	Matrix Input: Locate, read, store, tidy up; Read M, N, P	Requires args M (M), N (N), P (P);	13/04/2010 16/04/2010	Read and written, not stored. Stored in new ayb_model.
10	Processed Intensities: Calculate and store, tidy up	AYB struct taken from AYBc with int16 intensities replaced with tile	20/04/2010	
11	Initial Sequence: By maximum intensity	use call_base_simple	20/04/2010	
12	Sequence output: Create file, write data, close		21/04/2010	NUC needs raw type file? NUC changed to use XFILE
16	Matrix: Add read_methods for multiple styles	1) Intensities 2) As rows of columns. Move matrix read from cluster.	13/04/2010	New functions new_MAT_from_line, read_MAT_from_column_file
17	Initial Lambda: Ignore weights	use estimate_lambdaOLS	26/04/2010	

**Notes**

array.def  
tile  
dirio

Changes from git central  
Changes from git central  
Store location of predetermined matrices; new open\_matrix; new method returns name of current file (for message).

matrix

set\_MAT, transpose\_inplace, invert taken from AYBc.

lapack

For matrix, getrf/i taken from AYBc.

nuc

Taken from AYBc; defines NUC and PHREDCHAR types; isprob taken from utility.

New rcons\*\_list to appends to a given node (should be last)  
New read\_TILE keeps cluster list in input file order  
set\_path becomes more generic set\_location;

In invert change WORK/WORKSPACE type to real\_t

change WORK type to float in sgetri  
Read/show changed to use XFILE; show\_PHREDCHAR print space if out of range; replace printf with message  
NUC\_\*, \*\_PHRED do not need to be public? NUC\_ to enum?  
What about array construct?

call_bases	New; parts taken from AYBc call_bases (more later).	
intensities	New; parts taken from AYBc process intensities (more later).	
lambda	New; parts taken from AYBc estimate_lambda (more later).	
ayb_model	New; parts taken from AYBc ayb. (more later).	show_AYB use fp not stderr
utility.h	Remove NBASE def - now in nuc	
<b>Consider</b>		
ayb_model	Make bases & quals a list (per cluster)	
ayb_model	Does AYB struct need to be public at all? or even the standard functions?	

Index	Task	Additional Requirements	Completed	Notes
6	File I/O: Locate, open/create, read/write, close	Requires args input (i), output (o); search input dir for pattern matched files.	24/03/2010	
7	Data Structures: Tile, Cluster, Matrix	Use as is from Central Repository	01/04/2010	
8	Intensities Input: Locate, read, store, tidy up	Requires arg ncycles (n)	01/04/2010	
9	Matrix Input: Locate, read, store, tidy up; Read M, N, P	Use ncycle	postponed	
13	Replace message macros with vfprintf	Change call function name to message	12/03/2010	
14	Expand message severity list to include debug		24/03/2010	
15	Configure Doxygen; add comments to new HM files		29/03/2010	Adopted files still to do

Index	Task	Additional Requirements	Completed	Notes
1	Initial system: Make; Run	IDE develop but also build from command line	10/03/2010	Make with gcc. Makefile.
2	Program Version	Version file	08/03/2010	Store version and date.
3	Program arguments: Read and store (infrastructure)	Initial args help, licence, version, usage (default)	08/03/2010	Use getopt_long; Use include file method to do bulk output.
4	Program Log: Infrastructure	Log message from a Type and Severity; allow for parameters of varying type. Output to unique filename (from date/time) in configurable location.	10/03/2010	Requires args logfile (e), loglevel (l); Hide implementation from user.
5	Signal Handler	Initially interrupt and floating point exception	10/03/2010	On interrupt get confirmation first. Divide by zero does not cause FPE so not tested