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	
8	Quality score calibration, missing data handling, GC content, sim output, argument defaults, optimisation.	01/11/2010	12/11/2010	15/11/2010	1.07	
9	Additional final MPN output, documenting, tidy up.	15/11/2010	06/12/2010	07/12/2010	2.00	First public release
10	Quality calibration file and cif run-folder.	13/12/2010	21/01/2011	21/01/2011	2.01	
11	Quality calibration file now values to use.	01/02/2011	17/02/2011	17/02/2011	2.02	
12	Model refactor and automated testing.	22/02/2011	17/03/2011	10/05/2011	2.03	Delayed until needed
13	Memory use reduction and sim output changes	13/07/2011	22/07/2011	22/07/2011	2.04	
14	Improve modelling algorithms	26/07/2011	13/10/2011	18/10/2011	2.05	
15	Improve quality scoring	08/11/2011	24/11/2011	01/12/2011	2.06	Delayed due to sickness

16	Option to run with multiple threads	06/12/2011	16/12/2011	16/12/2011	2.07	
17	Use spike-in data, update system test	04/01/2012	06/02/2012	21/02/2012	2.08	Delayed due to incorrect estimate and new OS
18	Urgent bug fixes; improve handling of missing data.	04/04/2021	04/04/2012	04/04/2012	2.09	Released from branch v2_08_fixes

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, illlumina 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)
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.
46	Calibration of quality score.	8	8	Adjust in line with empirical observations using a table and neighbouring scores.
47	Deal with missing data by setting base call to ambiguous.	8	8	Defined as all intensities zero.
48	Additional deltas; use for M solve and add for covariance.	8	8	
49	Option to penalise base calls by genomic CG content.	8	8	Requires arg composition (c); values 0<gc<1.
50	Tidy compiler warnings from fortran and array.def.	8	8	
51	Optional output of full covariance matrix and lambda fit.	8	8	For multiple blocks use lambdas from first block only and append additional block covariance. Print header text, program version and command line, substituting for header text as already printed, then num cycles and fit params. All lines to begin with hash; locate any newlines and add hash. Requires arg simdata (s)
52	Final processed output does not allow for multiple blocks.	8	8	
53	Changes to arguments: default dataformat to cif, format to fastq, solver to zero. New blockstring default all available cycles in one block. Replace prefix with any number of non-option arguments, also allowing inclusion of partial path.	8	8	Requires additional input file search loop. Use of prefix has caused problems when e.g. file and file_end1 both exists. Make default prefix behaviour an exact match with a '+' as last character indicating treat as prefix. Message file must now be supplied as path and filename.
	Flush error file after each message to capture info in case of failure.	8	8	with 53
	Change message to use CSTRING instead of fixed length	8	8	with 53
54	Ensure an info message for all options that affect results.	8	8	Need GC comp and Mu; ignore message file. No genome composition if default.
55	Optimisation	8	8	Improve runtime.

56	Tidy compiler warnings from intel C compiler.	8	8	strcasecmp in strings, mode_t in sys/types, BSD_SOURCE for scandir, cast int to enum, goto jumps over variable declarations, multiple const
57	Additional output of M, P and N matrices.	9	9	With final processed, in correct format for input.
58	Does not seem to respond to ctrl<c>	9	9	
59	Use actual executable name in final message.	9	9	Useful for comparisons.
60	Phasing estimation bug.	9	9	
61	Fix warnings from static analyser.	9	9	
62	Expand and improve documentation.	9	9	Expand top level doxygen page; create manual page/user guide; improve and tidy up some module and code documentation.
63	Go to next prefix if input matrix is the wrong size.	9	9	Assume co-located intensities files with common prefix have same number of cycles.
64	Fix remaining compiler warnings.	9	9	xio discards qualifiers.
65	Create README with build instructions and outline CHANGELOG.	9	9	Use AsciiDoc format so suitable for web and download.
66	Read in quality calibration file. Use as conversion to default and optionally output result.	10	10	Format is column file with multiple matrices per file; includes comment lines and trailing comments (#). Requires arg qualtab (Q).
67	Read cif files directly from a run-folder.	10	10	Requires arg runfolder (r) (no argument), run-folder given in input option. Prefix replaced by lanetile string Ln[-n]Tn[-n] to process a range of lanes and tiles. Error if cif not selected. Output to s_L_TTTT. Note: single cycle fails with tmp memory error (MNP).
68	Allow prefix or input filepath to contain complete path.	10	10	May be more intuitive.

69	Make quality calibration file values to use instead of conversion for default. Need to co-ordinate release with new ayb_recal.	11	11	Need to output used values by default for use by ayb_recal, turn off with argument noqualout (q). One per run with filename logname.tab or ayb_xxxxxx_yymmdd_hhmm. Use header at top of table to act as history, separated from rest by a blank line.
70	Add success/failure message at end of log as well as to stdout.	11	11	with item 69
71	Split ayb_model which has become very large.	12	12	Extract bulk of AYB structure access to new ayb. Make pointer to AYB structure public but keep internals hidden.
72	Automate module testing with a test script.	12	12	Create reference results for future comparison. Enhancements to cluster/tile/nuc; expand message, xio to all; test all functions including null values and other bad input.
73	Automate system testing with a test script.	12	12	Create reference results for future comparison. Test normal run modes but not failure modes (done at module level). Test program options in principle.
74	Store intensities as integers.	13	13	Reduce memory use with little (if any) difference to results. Cif files integer anyway.
	In Matrix, test instructions are wrong way round	13	13	with 74
75	Changes to sim data output; coordinate with simNGS version 1.5 (12/05/11)	13	13	Lambda fit for each block; version no (5); option to vary fit distribution (weibull, normal, mixed normal, logistic) and selection char; fix as logistic.
76	Improve algorithms; new call bases, parameter A replaces M & P, new estimate lambda.	14	14	Improved results; more mapped and perfect.
	Set calls to null on data error.	14	14	with 76
77	Improve quality scoring.	15	15	Use similar algorithm to new call bases. New arg generr (g); replaces arg mu (m) in help etc.
78	New estimate lambda.	15	15	Previously not implemented as seemed to make results worse, but considerably improves results with some sets of data.
79	Minor changes to omega and phred calc.	15	15	To make more robust.

80	Simplify calculate covariance.	16	16	Remove now unused original (inner sum) to avoid need for mat array return.
81	Option to run with multiple threads.	16	16	Use OpenMP in selected functions (see mail 10/06/11). New arg parallel (p).
82	Use spike-in data to improve base calling and calibrate qualities.	17	17	see mail 08/09/11; New args spikein (K) and spikeuse (k); document new i/o.
83	Output not determinate when run multi-threaded (found in system testing)	17	17	Minor differences each run, mostly quality but occasionally a base.
84	AYB crashes if a negative number of iterations supplied.	17	17	Traced to ZeroLambda array failing to create.
85	System test for items 76 to 81.	17	17	Changes to options and major changes to outputs. Adjust developer doc.
86	Handle variance is zero in update cluster weights.	18	18	Can reset if no variance.
87	Fix problems in fit omega if cholesky fails.	18	18	Reset omega if factorisation fails.
88	Robust fixes to solve Nick's basecalling problem.	18	18	Better initial cross-talk; use cholesky based solver for linear system; memory corruption bug when intensities are missing.

In parameter calculation ignore clusters with too much missing data.	Count number of cycles with missing data. Include in parameter calc if less than new arg zerolimit (z)??.
Order of warning and information is incorrect in help and manual.	Not altered with change to log in item 31.
Tidy function headers for doxygen.	Weibull
Dirio and message program output not documented in header.	Do with next change.
Automated module testing continued. Plan: datablock, dirio, weibull?	Others to improve: matrix, mpn, Others new: ayb_options, cif, datablock, dirio, qual_table, utility, more? weibull?

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).
Additional for developer guide.	Overview, program flow, file formats (incl quality conversion), release instructions, glossary. Full stops at end of lists. Also how to generate in readme. Handling of array/list. Multi-threading.
Additional somewhere	Input file formats: matrix, quality table.
Investigate doxygen extension mapping warning.	Also expansion of macros; MACRO_EXPANSION=; def->h?
Web based bug tracking facility.	
Continue to look at efficiency.	

Errors

Functions that take no arguments should include void inside the argument brackets.	Deal with as seen; handler checksignals.
Doxy changes	Deal with as seen; model WORKPTR, datablock DATABLOCK

Consider

Remove dependency on message from nuc?	
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		
	Rationalise use of int/long/uint32 types	Be aware of efficiency issues.
	Optionally use message in matrix/cluster/tile/nuc/more? controlled by compiler switch.	Set switch flag in message?
	New message functionality to return message string	For use with perror, length of string? vsnprintf? Maybe return pointer to string stored and freed in message.
	Message: use a different word than type for msgtype?	Confusing in documentation
	Add segmentation violation handler	
	Tidy up order of some items, e.g. ayb_model set.., list of options.	
	Make --version option comply to GNU coding standards	
Note		
	matrix not yet documented	

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
	In parameter calculation ignore clusters with too much missing data.	Count number of cycles with missing data. Include in parameter calc if less than new arg zerolimit (z)??.		Ayb new boolptr missdata in ayb struct for missing data cluster list and global MissDataLim.

test with multi-thread

Index	Task	Additional Requirements	Completed	Notes
86	Handle variance is zero in update cluster weights	Can reset if no variance.	04/04/2012	Ayb update_cluster_weights reset weights to 1 if var is zero.
87	Fix problems in fit omega if cholesky fails.	Reset omega if factorisation fails.	04/04/2012	Matrix cholesky return null on error. Conjugate fit_omega reset omega if factorisation fails.
88	Robust fixes to solve Nick's basecalling problem.	Better initial cross-talk; use cholesky based solver for linear system; memory corruption bug when intensities are missing.	04/04/2012	Mpn solverChol add diagonal delta (like others). Ayb change initial_crosstalk; estimate_mpn replace solverSVD with solverChol; initialise remove different handling of missing data (i.e. do not set to 'N'; no longer set to 'N' after initialisation anyway).

Index	Task	Additional Requirements	Completed	Notes
82	Use spike-in data to improve base calling and calibrate qualities.	see mail 08/09/11; New args spikein (K) and spikeuse (k); document new i/o.	15/02/2012	New spikein to handle spikein data for single cluster and list; read from file and get next data for next cluster. Dirio new iotype spikein, add to set_location to store spikein path; new generic open_input_blk called by new open_spikein which also stores last filename returned by new get_last_spikein; new set_spikein determines actual path, called from startup; Call_bases new calculate_lss without calling bases. Nuc new max quality const for looping through qualities; new qualint_from_quality/phredchar; Ayb_model initialise requires block param; go to next block instead of exit loop if initialise fails. Ayb new boolptr spiked in ayb struct for spikein cluster list, new qspike struct for counts and new global flags; new calibrate_by_spikein can calibrate and replace or output count tables; new calibrate_by_table (moved); new read_spikein_data sets spiked and stores sequence in struct bases; in estimate_bases calc lss only for spikein clusters unless final iter; on last iter add spikein obs/diff to counts then call spikein calibrate after loop; in initialise read spikein data and skip call base if spikein; in output_final show effDF; qual calib message on startup.
82	(Module Test)	New spikein, message, nuc.	17/02/2012	Add spikein to module script. Message new messages and some with different params. Nuc new tests for new funcs and adjust test probability set to give more varied results.
83	Output not determinate when run multi-threaded (found in system testing)	Minor differences each run, mostly quality but occasionally a base.	16/01/2012	In ayb calculate_covariance the weight sum is accumulated into a single variable which may be done in a differing order when multi-threaded. Use a per-thread array instead and accumulate at end of parallel.
84	AYB crashes if a negative number of iterations supplied.	Traced to ZeroLambda array failing to create.	27/01/2012	strtoul can return a negative value which is then stored in an unsigned variable; when used in calloc it returns null. In ayb_model and ayb_options use strtol instead and check for negative.

85	System test for items 76 to 81.	Changes to options and major changes to outputs. Adjust developer doc.	17/01/2012	All reference output except calibration tables changed to match new algorithms; cif test file s_2_0001 replaced by larger s_2_0008 for correct full covar calc; gcs renamed to par; arguments c, m, M, P, S removed; arguments A, g, p added; check for fixed supplied A and N.

spike-in data

Aim: Modern chemistries add a small amount of a known genome, with an identifying tag, that can be used to improve the analysis of the data.

Function: Input a tab delimited file with cluster number and sequence on each line. Clusters contained in this file are set to sequence given in the file and not base-called. After the final iteration, use the extra information to calibrate the qualities instead of calibration file. Options to do full recalibration or just output numbers for whole run recalib.

Interfaces: Spike data file names will be {filename}[x].spike or {filename}[x]_spike.txt. Path given in option.
Diff and obs tables will be {filename}[x].qspike or {filename}[x]_qspike.txt

Design: Store a boolean array of known clusters. Initialise bases accordingly and then do not call.
In final iteration:
For known clusters:
1. Call each sequence and quality
2. For each predicted quality value, count number of different base calls and number in total
3. Calculate actual quality for each predicted quality (optional)

For all clusters, replace each predicted quality value with actual (including known clusters for which output called base) (optional); Or output diff and obs tables.

Algorithm: Qp = predicted quality, Nerrs = number different, Nobs = number total
Qa = actual quality (to find)
Pp = 10 ^ (-Qp/10)
Pa = (Nerrs + Pp)/(Nobs + 1)
Qa = -10 * log10(Pa)

Index	Task	Additional Requirements	Completed	Notes
80	Simplify calculate covariance.	Remove now unused original (inner sum) to avoid need for mat array return.	12/12/2011	Ayb remove accumulate_covariance; calculate_covariance only do full covar and all return values now a single mat; called from estimate_bases and ayb_model output_simdata.
81	Option to run with multiple threads.	Use OpenMP in selected functions. New arg parallel (p).	14/12/2011	Ayb_options new NThread and get/set; stored here because main has no associated header file for public functions. Ayb_main request number of threads, and report how many supplied. List.def new create/free an array of list pointers; allows use of parallel for loop; note ncluster must be unsigned int. Tile test array of list pointers. Ayb add multi-threading to calculate_covariance, estimate_bases & initialise; move show processed to outside of parallel (need to redo process intensities) for correct file output; amend error handling so no jumps out of parallel. Mpn add multi-threading to calculateNewJ/K.
81	(Module Test)	Message, Tile.	14/12/2011	Messagelog.ref new thread message. Tile.ref test array of list pointers.

Index	Task	Additional Requirements	Completed	Notes
77	Improve quality scoring.	Use similar algorithm to new call bases. New arg generr (g); replaces arg mu (m) in help etc.	25/11/2011	Call_bases new global generalised error val (PolyQual) with get/set; return value from call_bases; replace call_qualities with call_qualities_post. Message replace opt_select_se with opt_select_sg (g => f or e depending upon size plus no dp if zero). Statistics new median (and subs). Ayb new mat lss in struct; initialise to 4 * ncycle; in estimate_bases if last iter calculate effDF as median of lss; then update lss as return from call_bases; use replacement call_qualities_post, with effDF; in startup output gen err value instead of mu.
78	New estimate lambda.	Previously not implemented as seemed to make results worse, but considerably improves results with some sets of data.	21/11/2011	Lambda new estimate_lambda_A takes params raw intensities, At and N. Replace calls in ayb initialise and estimate_bases.
79	Minor changes to omega and phred calc.	To make more robust.	30/11/2011	Conjugate new const almost_one (match value to aybc) to use in transform/dtransform; in linemin also limit negative adjustment; in fit_omega add deltas to initialisation and increase max iterations to 400. Nuc in conversions from prob remove isprob check (allow to overflow and check later); in convert to phred if nan (p>1) set to min else if inf (p=1) then set to max.
77/79	(Module Test)	Message _se -> _sg and expand to test multiple number types. Phred calc changes; change test probability set to give better results.	30/11/2011	Messagelog.ref _se to _sg. Nuc.ref phred changes.

Index	Task	Additional Requirements	Completed	Notes
76	Improve algorithms; new call bases.	Improved results; more mapped and perfect.	21/09/2011	New conjugate calculates restricted fitted omega from full covariance using conjugate gradient method. Lapack new getrs (intensities), potri (matrix), trmm (conjugate), syr (ayb). Matrix new invert_symmetric for sym pos def mat (not used); cholesky improved. Call_bases new call_bases (and subs) calls every base in a cluster using dynamic programming algorithm; new call_qualities with original algorithm but using pre-called bases. Nuc show_phredchar change default if unprintable to min because bwa does not recognise space. Ayb add omega to struct; in accumulate_all_covariance use lapack syr to calculate; in estimate_bases calculate full covariance and get cycle_var from and fitted omega, (covar as array no longer needed but keep), use new call_bases and call_qualities; only call qualities if last iter (not needed before); do not re-estimate lambda if last iter so good for final values.
76	Improve algorithms; parameter A replaces M & P.	A stored & used as A transpose. New argument A (A) replaces P (P); arguments composition (c) and solver (S) no longer used. If A and N supplied (both or neither required) then keep them fixed.	11/10/2011	Dirio ParamA replaces Phasing in typedef. Intensities new ProcessNew calculates proc int using At instead of M & P. Matrix new structLU for At and piv and LUdecomposition. Mpn new calculate_NewJ/NewK/Lhs/Rhs. Ayb add At to struct (remove M & P ??) new param solver const; new globals FixedParam and Initial_At initialised to blocks of initial M in init_matrix; replace M & P with At in update_cluster_weights, output_final, calculate_covariance, estimate_bases, estimate_MPN (NewJ etc), initialise_model; also replace process_intensities (and expected_intensities) with processNew; estimate_MPN now much simpler with no need to iterate, skip estimation loop if FixedParam; in initialise_model M now temporary and store A as transpose and initial; in startup check both or neither of options A & N supplied and set FixedParam; remove all references to composition (base penalty) and solver.

				Options remove composition (c) and solver (S); replace P with A.
76	new estimate lambda	Not implemented as made results worse.		
76	Check null returns	ProcessNew, fit_omega	12/10/2011	Ayb new set_null_calls, call whenever processing fails; check for null return from ProcessNew, fit_omega; also check for error in estimate_MPN and return nan. Ayb_model analyse_tile check return from estimate_MPN, stop if nan.
76	Reduce calls to ProcessNew.	Call to ProcessNew for update weights can be removed by storing least squares error after base call.	12/10/2011	Ayb new store_cluster_error containing error calc for a single cluster extracted from update_cluster_weights and remove cluster loop and call to ProcessNew; call from within cluster loop in initialise_model and estimate_bases after lambda calc; do not call if last iteration so final weights value saved.
76	(Module Test)	Unprintable phredchar output change to min. New err message for A & N matrix supply.	13/10/2011	Nuc change test wording and new ref. New messagelog ref.
76	Remove fortran dependency.	NNLS no longer used.	14/10/2011	Mpn surround solverNNLS with compiler switch (FORTRAN). Makefile remove s/dnnls.o from object list and fortran library from flags. nnls references left in lapack and mpn.h.
76	Bad data causes overflow in omega fit.	Add some minor adjustments to algorithm.	10/10/2011	Conjugate transform add a near one multiplier to avoid an equality; linemin_obj limit the adjustment.
76	(System Test)	Changes to options and major changes to outputs.		Not yet done

Index	Task	Additional Requirements	Completed	Notes
74	Store intensities as integers.	Reduce memory use with little (if any) difference to results. Cif files integer anyway.	24/05/2011	Matrix add alternative integer array to structure and flag to indicate; use defined int_t with defined limits and print format; new clipint ensures integer value within range; new new_MAT_int takes type parameter (call from new_MAT with real default); add to free, copy, show and trim; add to append with type mismatch check; in new_MAT_from line create an integer matrix; new coerce from intarray (original is real). Cif new cif_get_int. Cluster make signals matrix int_t; affects coerce from array and read from cif; also ayb nodata, cycle_ints; also process_intensities (and add I-N precalc); also mpn calculate Ibar and K; also tile coerce from array.
74	(Test Ref)	Results from txt file input will change.	24/05/2011	Module: matrix append type mismatch (also improve append row mismatch) and append int type; cluster and tile show cluster values as int; mpn show int values and test values x 100 to compensate; msgerr append type mismatch err msg. System: txt and zip output.
75	Changes to sim data output; coordinate with simNGS version 1.5 (12/05/11)	Lambda fit for each block; version no (5); option to vary fit distribution (weibull, normal, mixed normal, logistic) and selection char; fix as logistic.	19/07/2011	Ayb_model output_simdata move lambda fit calc and output to outside of first block condition; add sim version 5; introduce distribution selection but fix at logistic; add selection char to output. New mixnormal has routines for fitting a mixed normal distribution.
75	(Test Ref)	Mixnormal has module testing.	19/07/2011	Module: add mixnormal to module script with new input and ref files. System: add sim output to blk test; also txt and cif runfile output change.

Index	Task	Additional Requirements	Completed	Notes
71	Split ayb_model which has become very large.	Extract bulk of AYB structure access to new ayb. Make pointer to AYB structure public but keep internals hidden.	22/03/2011	See AYB model split plan below. Model keeps setup, read intensities, looping structure and output (except final). AYB structure pointer no longer global but passed as parameter.
72	Automate module testing with a test script.	Create reference results for future comparison. Enhancements to cluster/tile/nuc; expand message, xio to all; test all functions including null values and other bad input.	16/03/2011	Cif readCIFfromStream check for null fp and read file failure. Cluster show increase maximum matrix columns to 10; read_first_cluster, all cycles, write_coordinates, optional cif input file testing. Message expand to test all functions; in message lists add dummy enums between different param requirements and use to bulk test. Nuc nucs_from_string, quality range, quality from prob, optional read_nuc/phred testing. Tile read_folder_tile check for null root and store lane tile; all cycles, write_lane_tile, copy_tile, optional cif input file and folder testing. Xio expand to test all functions; xfgettok check for null separator; include bad parameter and bad input testing; corrections made: add extra checks for null parameters; xfgets returns null if fails; xfgettok zero len returned if bad parameter; correct return type from zip read/write; use fputs not xfputs (under test) for test result output. New AYB_module_test.sh, standard inputs and outputs.
73	Automate system testing with a test script.	Create reference results for future comparison. Test normal run modes but not failure modes (done at module level). Test program options in principle.	06/04/2011	New AYB_system_test.sh, standard inputs and outputs in a suite of directories.

Ayb_model split plan

AYB structure referenced in:

function	called by	uses	destination	new structure access functions
initialise_model	analyse_tile	All	ayb (public)	
update_cluster_weights	estimate_MPN	All	ayb	
estimate_MPN	analyse_tile	All	ayb (public)	
calculate_covariance	estimate_bases, output_simdata	All	ayb (public)	
open_processed	estimate_bases	ncycle, ncluster	ayb	
write_processed	estimate_bases	ncycle, tile	ayb	
output_final	close_processed	whole, M, N, P	ayb	
close_processed	estimate_bases	whole	ayb	
estimate_bases	analyse_tile	All	ayb (public)	
output_results	analyse_tile	whole, ncluster, ncycle, bases, quals	model	get_ncluster, show_bases, show_quals
output_simdata	analyse_tile	lambda, ncycle	model	get_ncycle, get_lambdas (nonzero)
public	----------	----------	----------	----------
new_AYB	analyse_tile	All	ayb	
free_AYB	analyse_tile	All	ayb	
copy_AYB	-	All	ayb	
show_AYB	analyse_tile, estimate_bases, output_final	All	ayb	
analyse_tile	ayb_main	whole, tile, ncycle	model	get_ncycle, replace_tile

AYB structure not used in:

function	called by		destination	need
read_matrices	startup		ayb	call to ayb_startup
create_datablocks	analyse_tile		model	
init_matrix	initialise_model		ayb	
nodata	initialise_model, estimate_bases		ayb	
accumulate_all_ covariance	calculate_covariance		ayb	
accumulate_covariance	calculate_covariance		ayb	
output_zero_lambdas	analyse_tile		model	
has_whitespace	output_simdata		model	
format_header	output_simdata		model	
public	----------	----------	----------	----------
read_intensities_file	ayb_main		model	
read_intensities_folder	ayb_main		model	
startup_model	ayb_main		model	
tidyup_model	ayb_main		model	

option function	used in	sets	destination	need
set_composition	startup, estimate_bases	BasePenalty	ayb	log message also
set_niter	startup, analyse_tile, output_zero_lambdas	NIter	model	
set_output_format	startup, output_results	OutputFormat	model	
set_show_working	estimate_bases	ShowWorking	ayb	
set_simdata	analyse_tile, output_simdata, tidyup	SimData, SimText	model	
set_solver	startup, estimate_MPN	SolverIndex, SolverRoutine	ayb	log message also

Move to ayb_startup:

some log messages
read_matrices
read_quality_table

Constants and members moved as required.

Index	Task	Additional Requirements	Completed	Notes
69	Make quality calibration file values to use instead of conversion for default. Need to co-ordinate release with new ayb_recal.	Need to output used values by default for use by ayb_recal, turn off with argument noqualout (q). One per run with filename logname.tab or ayb_xxxxxx_yymmdd_hhmm. Use header at top of table to act as history, separated from rest by a blank line.	10/02/2011	New qual_table contains quality table object with structure shared with ayb_recal (same filename); read in, output and adjust_quality (from call_bases); home of qualout flag; default table stored in new calibration where header contains just external Qtable pointer and source generated by ayb_recal. Xio new xfgettok read until separator; xfgetln use with separator \n; return null instead of empty string if nothing read, requires null checks in cluster; new test-xio, also include in Makefile. Message new generate_name creates virtual filename using dev/urandom; call from startup_message if not specified; new get_message_path; new test-message, also include in Makefile. Dirio new open_run_output makes new output filepath from log file and opens; uses new name_only and renamed output_name_suffix. Call_bases all quality table moved to qual_table, affects ayb_model. Ayb_main call output_quality_table, tidyup_qual_table Ayb_options opt simdata moved to first in longopts list to avoid having to continually change simdata index.
70	Add success/failure message at end of log as well as to stdout.	with item 69	10/02/2011	

Index	Task	Additional Requirements	Completed	Notes
66	Read in quality calibration file. Use as conversion to default and optionally output result.	Format is column file with multiple matrices per file; includes comment lines and trailing comments (#). Requires arg qualtab (Q).	17/12/2010	Matrix read_column_file already ignores anything trailing after sufficient column values but check each value for valid numeric; skip comment lines using new getnextline; write_column_file optional free format outputs values with default decimal places. Dirio store filepath in same array as MNP, need new end MNP in enum; Call_bases new read_quality_table (using 3 new subproc); output if msg level debug; call from startup_model. Message new get message level.
67	Read cif files directly from a run-folder.	Requires arg runfolder (r) (no argument), run-folder given in input option. Prefix replaced by lanetile string Ln[-n]Tn[-n] to process a range of lanes and tiles. Error if cif not selected. Output to s_L_TTTT. Note: single cycle fails with tmp memory error (MNP).	20/01/2011	Tile new read_folder_TILE; move common cif to tile to new create_TILE_from_cif. Dirio new int pair type for lane/tile; new globals for min, max and current; runfolder flag set by option; new check_dir checks a dir exists, extracted from check_outdir; replace get_pattern (no longer used) with get_input_path; new set_lanetile (and 3 new subproc) decodes lanetile string and stores, call from set_pattern if runfolder; new get_next_lanetile returns next lane/tile and sets current filename; also lanetile_isnull; in startup check input dir exists and if runfolder check cif selected. Datablock use message log. Cif errors: crash on large values found in cif_set_from_real switch; use new round_and_clip; file not closed in cif_add_file; also pass cluster/cycle to showCIF. Ayb_model make read_intensities_file global and new read_intensities_folder; store result in new global MainTile and move error checks to analyse_tile, removing read_intensities call; check at least 2 cycles. Ayb_main if runfolder call get_next_lane_tile instead of open_next; call read_intensities_file/folder before analyse_tile.
68	Allow prefix or input filepath to contain complete path.	May be more intuitive.	07/01/2011	Dirio full_path and move_path (was move_partial_path) check filename for root dir and use alone if found.

Index	Task	Additional Requirements	Completed	Notes
57	Additional output of M, P and N matrices.	With final processed, in correct format for input.	16/11/2010	New matrix write_MAT_to_column_file. Ayb_model new output_final.
58	Does not seem to respond to ctrl<c>		16/11/2010	No tidyup required so just remove handler install from main.
59	Use actual executable name in final message.	Useful for comparisons.	16/11/2010	ayb_main from argv[0]. Also make path delimiter a common constant in dirio (also used in message).
60	Phasing estimation bug.		23/11/2010	ayb_model and mpn calculatePrhs.
61	Fix warnings from static analyser.		26/11/2010	Ayb_model unused return values, initialise_model, calc_covariance, estimate_bases cleanup; Cluster, dirio unused return values. Mpn default mode value. Utility incorrect extend_cstring, change to renew. Xio unused strlen.
62	Expand and improve documentation.	Expand top level doxygen page; create manual page/user guide; improve and tidy up some module and code documentation.	07/12/2010	mainpage, Doxyfile (also remove usefloat); new AYB.1.txt in AsciiDoc format, generates user and man page; Makefile generate user/man with archive; ayb_help, ayb_usage, datablock, dirio, message (also move fflush).
63	Go to next prefix if input matrix is the wrong size.	Assume co-located intensities files with common prefix have same number of cycles.	01/12/2010	Status fail on ayb_model false return from initialise_model.
64	Fix remaining compiler warnings.	xio discards qualifiers.	02/12/2010	Use a cast to remove.
65	Create README with build instructions and outline CHANGELOG.	Use AsciiDoc format so suitable for web and download.	07/12/2010	

Index	Task	Additional Requirements	Completed	Notes
46	Calibration of quality score.	Adjust in line with empirical observations using a table and neighbouring scores.	12/10/2010	Call base returns quality value not phred char; new adjust quality called from ayb_model after all bases called and before convert to phred; first and last require special handling; include new calibration files. New nuc routines to split prob to phred and MIN_QUALITY.
47	Deal with missing data by setting base call to ambiguous.	Defined as all intensities zero.	12/10/2010	Nuc ambig is beyond nbase array range; new nuc isambig called wherever base index referenced. New call_base_nodata; add handling to quality adjust. In ayb_model check for no data before call base.
48	Additional deltas; use for M solve and add for covariance.		12/10/2010	
49	Option to penalise base calls by genomic CG content.	Requires arg composition (c); values 0<gc<1.	12/10/2010	Penalty calculated for each base and stored as arrayx4. Pass to call base and use to adjust calculation.
50	Tidy compiler warnings from fortran and array.def.		12/10/2010	Array use attribute used.
51	Optional output of full covariance matrix and lambda fit.	For multiple blocks use lambdas from first block only and append additional block covariance. Print header text, program version and command line, substituting for header text as already printed, then num cycles and fit params. All lines to begin with hash; locate any newlines and add hash. Requires arg simdata (s)	22/10/2010	Dirio open append option. Matrix show_MAT with optional rownum. Ayb_options match string to an option. Version now in new ayb_version.c so version can be exported to more than one file. Ayb_model store flag and header text; accumulate_all_covariance; output_simdata calculates values and writes after formatting header lines; analyse_tile needs arg params from ayb_main.
52	Final processed output does not allow for multiple blocks.		22/10/2010	ayb_model pass blk ->estimate_bases -> open_processed; use to open pif and final.

53a	Changes to arguments: default dataformat to cif, format to fastq, solver to zero.		27/10/2010	dataformat - dirio change default; ayb_model remove duplicate InputFormat - get when needed; info message moved to dirio (improve order). output format - ayb_model change default. solver - ayb_model change default; store solver index and output info message; make selection info message generic. Update help defaults.
53b	New blockstring default all available cycles in one block.		28/10/2010	
53c	Replace prefix with any number of non-option arguments, also allowing inclusion of partial path.	Requires additional input file search loop. Use of prefix has caused problems when e.g. file and file_end1 both exists. Make default prefix behaviour an exact match with a '+' as last character indicating treat as prefix. Message file must now be supplied as path and filename.	09/11/2010	Dirio move pattern check/scandir to set_pattern; new clear_pattern; new pattern path, filled by new move_partial_path, to hold input path and partial path; in match_pattern treat as prefix only if prefix indicator. Read_options return next arg index; options/help/usage make alphabetical. Move retopt enum to utility and use as analyse_tile return to indicate stop/next pattern; propagate in ayb_model as required. Ayb_main new outer loop through non-option args calling set_pattern. Message path is now full pathname not location and no default; create_filename replaced by check_message_path; some fatal messages now error.
54	Ensure an info message for all options that affect results.	Need GC comp and Mu; ignore message file. No genome composition if default.	02/11/2010	Ayb_model output from model startup; mu needs export from call_bases.
55	Optimisation	Improve runtime.	11/11/2010	Nuc new has_ambiguous_base used as precheck where isambig is called in tight loop; replace int/uint_32 with uint_fast32 in compute heavy loops; precalculate weight matrix; affects intensities process/expected_intensities, mpn calculateJ/K. Makefile add unroll-loops optimiser.
56	Tidy compiler warnings from intel C compiler.	strcasecmp in strings, mode_t in sys/types, BSD_SOURCE for scandir, cast int to enum, goto jumps over variable declarations, multiple const	12/11/2010	ayb_main, ayb_model, cif, cluster, dirio, tile, utility, Makefile

Index	Task	Additional Requirements	Completed	Notes
38	Optional output of final processed intensities, M, P, N, lambda, weights	Final processed format as intensities input, illlumina 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 (NFORTRAN) 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	Changes from git central	New rcons*_list to appends to a given node (should be last)
tile	Changes from git central	New read_TILE keeps cluster list in input file order
dirio	Store location of predetermined matrices; new open_matrix; new method returns name of current file (for message).	set_path becomes more generic set_location;
matrix	set_MAT, transpose_inplace, invert taken from AYBc.	In invert change WORK/WORKSPACE type to real_t
lapack	For matrix, getrf/i taken from AYBc.	change WORK type to float in sgetri
nuc	Taken from AYBc; defines NUC and PHREDCHAR types; isprob taken from utility.	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	

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?	

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