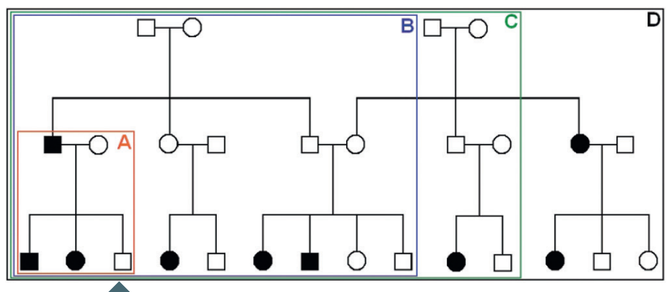
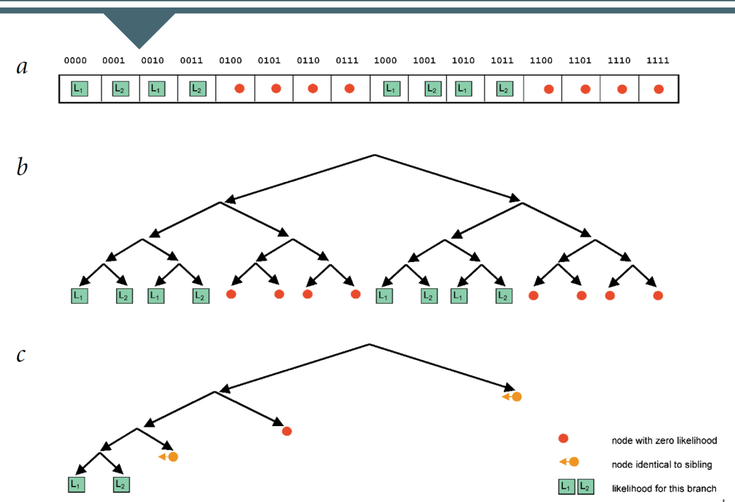
The Wellcome Trust Centre for
Human Genetics

UNIVERSITY OF OXFORD

Merlin--rapid analysis of dense genetic maps using sparse gene flow trees



Sample pedigrees used in simulations. Pedigree A is typical of affected sib-pair studies. Pedigree B, C and D are larger pedigrees used for benchmarking. Complexity (2n-f) is 4, 19, 25 and 32 bits for pedigrees A, B, C and D, respectively. If grandparents are not genotyped and male and female recombination fractions are assumed equal, complexity becomes 4, 18, 23 and 30 bits, respectively



Alternative representatives of gene flow in a pedigree. a, A bit-indexed array. The most common representation, this uses a sequence of binary digits, or an inheritance vector, to specify the outcome of each meiosis. Each of these sequences serves as an index into an array where the statistic of interest is stored. b, A packed tree in which individual meiosis are represented as new levels, and likelihoods or other statistics are evaluated for each leaf node. c, A sparse tree, in which each branch (meiosis) is evaluated conditional on the outcome of preceding meiosis. Evaluation stops early in sections producing invariant outcomes, resulting in premature leaf nodes (red circles). These occur, for example, when an impossible gene flow pattern is detected. Uninformative meiosis produce symmetric nodes and further increase sparseness (orange circle with arrow). These occur, for example, when both parental alleles are indistinguishable

WHAT WAS KNOWN

- The inheritance of disease-causing genes can be mapped to specific regions on a chromosome by studying families with multiple affected members in a linkage analysis
- This research approach has underpinned the identification of diseases caused by mutations in single genes (e.g. Huntington's disease) as well as diseases with more complex genetic architectures (e.g. type 1 diabetes)
- Technical advances in high-throughput genotyping could now generate very dense maps containing thousands of SNPs. Such maps had the potential to unambiguously track disease genes in families available to researchers, a benefit that would accelerate the discovery of disease genes
- Existing analytic algorithms were struggling to cope with this new dense genotype data even with access to high-performance computers. Accordingly, researchers were frustrated that important mapping information was inaccessible to them, delaying a comprehensive analysis of their expensively assembled data

WHAT WE DID

- We recognized that algorithms to model gene flows in families could be simplified by the adoption of sparse binary trees. This substantially reduced the complexity of the likelihood calculations that are required for linkage analysis, making the whole process much more rapid, as well as reducing the concomitant memory requirements
- An approximation to the algorithm that ignored extremely unlikely gene flow events, led to even more dramatic speedups in the calculations
- We integrated these improvements into the MERLIN software package, which could rapidly extract the maximal linkage information from precious family data. It was also able to identify implausible as well as erroneous genotypes by phasing haplotypes in families

WHAT THIS ADDS

- MERLIN was able to efficiently calculate linkage statistics tuned for the analysis of monogenic or polygenic diseases or quantitative inherited traits from extremely dense genotype data. Moreover, it was able to perform these analyses rapidly on affordable desktop computers
- The genotype error detection facility allowed researchers to eliminate potential noise in their data to increase power to detect true genetic signals. They were also able to assess if investment in additional genotyping would yield a useful increment in mapping information, to guide the design of further experiments
- MERLIN provided a one-stop-shop for linkage analysis, it was widely adopted by the gene mapping community and was instrumental in mapping numerous novel genes for inherited diseases. Examples include autism (*Nature Genetics* 2007 39, 319–328), the ALK gene in familial neuroblastoma (*Nature* 2008 455, 930–935) and the VCP gene in amyotrophic lateral sclerosis (2010 68, 857–64)

REFERENCES

Merlin--rapid analysis of dense genetic maps using sparse gene flow trees.

G R Abecasis, S S Cherny, W O Cookson & L R Cardon.

Nature Genetics 2001, 30, 97–101.