Supplementary Methods

Genomic DNA isolation and PFG analysis. High molecular weight DNA was purified from cell lines GM12877, GM12878, GM12885, GM12886, GM12891, GM12892 GM19240, and GM20431 (Coriell Institute for Medical Research, Camden, NJ) using a RecoverEase DNA isolation kit (Agilent, La Jolla, CA) following the manufacturer's protocol. High molecular weight DNA was partially sheared to make it more amenable to manipulation by pipetting 20-40 times using a Rainin P1000 pipette. 200 ng of genomic DNA was analyzed on 1% agarose gel with 0.5X TBE buffer using a BioRad CHEF-DR II with the following parameters: 6V/cm, 50-90 second ramped switch time, and a 20 hour total run. 500 ng of Yeast Chromosome PFG Marker (New England Biolabs, Ipswich, MA) and Lambda Ladder PFG Marker (New England Biolabs, Ipswich, MA) were used to determine the length of purified genomic DNA.

Isolation of 10 cells of NA19240. Immortalized cell line GM19240 (Coriell Institute for Medical Research, Camden, NJ) was grown in RPMI supplemented with 10% FBS under standard environmental conditions for cell culture. Individual cells were isolated under 200x magnification with a micromanipulator (Eppendorf, Hamburg, Germany) and deposited into a 1.5 ml microtube with 10 ul of dH_2O. The cells were denatured with 1 ul of 20mM KOH and 0.5 mM EDTA. The denatured cells were then entered into the LFR process.

Long Fragment Read technology. DNA was diluted and denatured at a concentration of 50 pg/ul in a solution of 20mM KOH and 0.5 mM EDTA. After a 1 minute incubation at room temperature 120 pg of denatured DNA was removed and added to 32 ul of 1 mM 3' thio protected random octamers (IDT, Coralville, IA). After two minutes the mixture was brought to a volume of 400 ul with dH_2O and 1 ul was distributed to each well of a 384 well plate. 1 ul of a 2X phi29 polymerase (Enzymatics Inc., Beverly, MA) based multiple displacement amplification (MDA)[22] mix was added to each well to generate approximately 3-10 nanograms of DNA (10,000-25,000 fold amplification). The MDA reaction consisted of 50 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 10 mM $(NH_4)_2SO_4$, 4 mM DTT, 250 uM dNTPs (USB, Cleveland, OH), 10 uM 2'-deoxyuridine 5'-triphosphate (dUTP) (USB, Cleveland, OH), and 0.25 units of phi29 polymerase.

Controlled Random Enzymatic Fragmentation (CoRE). Excess nucleotides were inactivated and uracil bases were removed by a 120 minute incubation of the MDA reaction with a mixture of 0.031 units of shrimp alkaline phosphatase (SAP) (USB, Cleveland, OH), 0.039 units of uracil DNA glycosylase (New England Biolabs, Ipswich, MA) and 0.078 units of endonuclease IV (New England Biolabs, Ipswich, MA) at 37°C. SAP was heat inactivated at 65°C for 15 minutes. A 60 minute room temperature nick translation with 0.1 units of *E. coli* DNA polymerase 1 (New England Biolabs, Ipswich, MA) in the same buffer with the addition of 0.1 nanomoles of dNTPs (USB, Cleveland, OH) resolved the gaps and fragmented the DNA to 300-1,300 base pair fragments. *E. coli* DNA polymerase 1 was heat inactivated at 65°C for 10 minutes. Remaining 5' phosphates were removed by incubation with 0.031 units of SAP (USB, Cleveland, OH) for 60 minutes at 37°C. SAP was heat inactivated at 65°C for 15 minutes.

Tagged adapter ligation and nick translation. Ten base DNA barcode adapters, unique for each well, were attached to the fragmented DNA using a two part directional ligation approach. Approximately 0.03 pmol of fragmented MDA product were incubated for 4 hours at room temperature in a reaction containing 50mM Tris-HCl (pH

7.8), 2.5% PEG 8000, 10mM MgCl2, 1mM rATP, a 100-fold molar excess of 5'-phosphorylated (5'PO4) and 3' dideoxy terminated (3'dd) common Ad1 (Supplementary Methods Figure 2) and 75 units of T4 DNA ligase (Enzymatics, Beverly, MA) in a total volume of 7 ul. Ad1 contained a common overhang region for hybridization and ligation to a unique barcode adapter. After 4 hours, a 200 fold molar excess of unique 5' phosphorylated tagged adapters were added to each well and allowed to incubate 16 hours (Supplementary Methods Figure 2). The 384 wells were combined to a total volume of ~ 2.5 ml and purified by the addition of 2.5 ml of AMPure beads (Beckman-Coulter, Brea, CA). 1 round of PCR was performed to create a molecule with a 5' adapter and tag on one side and a 3' blunt end on the other side. The 3' adapter was added in a ligation reaction similar to the 5' adapter as described above. To seal nicks created by the ligation, the DNA was incubated for 5 minutes at 60°C in a reaction containing 0.33 uM Ad1 PCR1 primers, 10mM Tris-HCl (pH 78.3), 50 mM KCl, 1.5 mM MgCl2, 1 mM rATP, 100 uM dNTPs, to exchange 3' dideoxy terminated Ad1 oligos with 3'OH terminated Ad1 PCR1 primers. The reaction was then cooled to 37°C and, after addition of 90 units of Taq DNA polymerase (New England Biolabs, Ipswich, MA) and 21600 units of T4 DNA ligase, was incubated a further 30 minutes at 37°C, to create functional 5'PO4 gDNA termini by Taq-catalyzed nick translation from Ad1 PCR1 primer 3' OH termini, and to seal the resulting repaired nicks by T4 DNA ligation. At this point the material was incorporated into the standard Complete Genomics' DNA nanoarray sequencing process[11].

RNA sequencing. The RNA-Seq data were derived starting from the total RNA, using the Ovation RNA-Seq kit (NuGen, San Carlos, CA) and SPRIWork (Beckman-Coulter, Brea, CA) to prepare a sequencing library with an average insert size of 150-200 bp. A 75bp paired-end sequencing reaction was performed on HiSeq 2000 (Illumina, San Diego, CA) at the Center for Personalized Genetic Medicine (Harvard Medical School, Boston, MA). Paired-end reads were assembled with tophat v1.2.0 using bowtie v0.12.7, and single nucleotide variants (SNVs) were called using the GATK UnifiedGenotyper v1.1 with hg19 for reference and dbSNP version 132 to annotate known SNPs. SNVs were mapped both to genes from RefSeq and to isoforms in the transcriptome as identified by cufflinks v1.0.3.

To identify haplotypes of co-expressed alleles, the data were filtered for heterozygous SNVs that occur both on the same LFR contig and on the same gene with at least one other heterozygous SNV. Where transcripts exhibit allele-specific expression, heterozygous alleles expressed on an LFR-phased haplotype should all have higher, or all have lower read counts than their counterparts on the other haplotype. Here we identify the higher-expressed haplotype as the one for which the majority of its het alleles exhibit higher expression than their counterparts. A heterozygous is counted as "concordant" if its expression agrees with its containing haplotype. In cases of ties, where there is no haplotype majority, half of the heterozygous SNVs are counted as concordant. Additionally, in order to be considered at all, the heterozygous SNV is required to have at least 20 fold RNA-Seq read coverage. The heterozygous SNVs are further filtered for noise from the GATK genotyper by comparing with the probability of choosing the ASE and coverage at random using the binomial test.

Reed-Solomon Barcode Error Correction

Each DNA nanoball (DNB)[11] is tagged with a 10-base Reed-Solomon code with 1-base error correction capability for the unknown error location, or 2-base error correction capability for when the errors positions are known. These 384 codes were selected from a comprehensive set of 4096 Reed-Solomon codes with the above properties. Each code from this set has a minimum Hamming distance of 3 to any other code in the set. For this study, the position of the errors is assumed to be unknown.

Basecalling Algorithm. Basecalling of the four images for each field is done in several steps. First, the image intensities are corrected for background using modified morphological "image open" operation. Since the locations of our DNBs line up with the pixel locations the intensity extraction is done as a simple read-out of pixel intensities from the background corrected images. These intensities are then corrected for several sources of both optical and biological signal cross-talks, as described below. The corrected intensities are then passed to a probabilistic model that ultimately produces for each DNB a set of four probabilities of the four possible basecall outcomes. Several metrics are then combined to compute the basecall score using pre-fitted logistic regression.

Intensity correction:

Several sources of biological and optical cross-talks are corrected using linear regression model. The linear regression was preferred over de-convolution methods that are computationally more expensive and produced results with similar quality. The sources of optical cross-talks include filter band overlaps between the four fluorescent dye spectra, and the lateral cross-talks between neighboring DNBs due to light diffraction at their close proximities. The biological sources of cross-talks include incomplete wash of previous cycle, probe synthesis errors and probe “slipping” contaminating signals of neighboring positions, incomplete anchor extension when interrogating “outer” (more distant) bases from anchors. The linear regression is used to determine the part of DNB intensities that can be predicted using intensities of either neighboring DNBs or intensities from previous cycle or other DNB positions (see equation below). The part of the intensities that can be explained by these sources of cross-talk is then subtracted from the original extracted intensities. To determine the regression coefficients the intensities on the left side of the linear regression model need to be composed primarily of only “background” intensities – i.e. intensities of DNBs that would not be called the given base for which the regression is being performed. This requires pre-calling step that is done using the original intensities. Once the DNBs that do not have a particular basecall (with reasonable confidence) are selected we perform a simultaneous regression of the cross-talk sources:

$$I_{background}^{Base} \approx I_{DNBneighbor1}^{Base} + \ldots + I_{DNBneighborN}^{Base} + I_{DNB}^{Base2} + I_{DNB}^{Base3} + I_{DNB}^{Base4} + I_{DNBpreviousCycle}^{Base} + I_{DNBotherPosition1}^{Base} + \ldots + I_{DNBotherPositionN}^{Base} + \varepsilon$$

The neighbor DNB cross-talk is corrected both using the above regression and also each DNB is corrected for its particular neighborhood using linear model involving all neighbors over all available DNB positions.

Basecall probabilities:

Calling bases using maximum intensity does not account for the different shapes of background intensity distributions of the four bases. To address such possible differences we have developed a probabilistic model based on empirical probability distributions of the background intensities. Once the intensities are corrected we pre-call some DNBs using maximum intensities (DNBs that pass certain confidence threshold) and use these pre-called DNBs to derive the background intensity distributions (distributions of intensities of DNBs that are not called a given base). Upon obtaining such distributions we can compute for each DNB a tail probability under that distribution that describes the empirical probability of the intensity being background intensity. Therefore, for each DNB and each of the four intensities we can obtain their probabilities of being background (p_{BG}^{A} , p_{BG}^{C} , p_{BG}^{G} , p_{BG}^{T}). Then we can compute the probabilities of all possible basecall outcomes using these probabilities. The possible basecall outcomes need to describe also spots that can be double or in general multiple-occupied or not occupied by a DNB. Combining the computed probabilities with their prior probabilities (lower prior for multiple-occupied or empty spots) gives rise to the probabilities of the 16 possible outcomes:

$$p^{A} = \frac{!\, p_{BG}^{A} + p_{BG}^{C} + p_{BG}^{G} + p_{BG}^{T}}{\sum p} * p_{SingleBase}^{prior}$$

$$p^{AC} = \frac{!\, p_{BG}^{A} + !\, p_{BG}^{C} + p_{BG}^{G} + p_{BG}^{T}}{\sum p} * p_{DoubleOccupied}^{prior}$$

$$p^{ACG} = \frac{!\, p_{BG}^{A} + !\, p_{BG}^{C} + !\, p_{BG}^{G} + p_{BG}^{T}}{\sum p} * p_{TripleOccupied}^{prior}$$

$$p^{ACGT} = \frac{!\, p_{BG}^{A} + !\, p_{BG}^{C} + !\, p_{BG}^{G} + !\, p_{BG}^{T}}{\sum p} * p_{QuadrupleOccupied}^{prior}$$

$$p^{N} = \frac{p_{BG}^{A} + p_{BG}^{C} + p_{BG}^{G} + p_{BG}^{T}}{\sum p} * p_{EmptySpot}^{prior}$$

These 16 probabilities can then be combined to obtain a reduced set of 4 probabilities for the four possible basecalls. I.e:

$$p_{4base}^{A} = p^{A} + \tfrac{1}{2}\left(p^{AC} + p^{AG} + p^{AT}\right) + \tfrac{1}{3}\left(p^{ACG} + p^{ACT} + p^{AGT}\right) + \tfrac{1}{4}\,p^{ACGT} + \tfrac{1}{4}\,p^{N}$$

Score computation:

Logistic regression was used to derive the score computation formula. The logistic regression was fitted to mapping outcomes of our basecalls using several metrics as inputs. The metrics included probability ratio between the called base and the next highest base, called base intensity, indicator variable of the basecall identity, and metrics describing the overall clustering quality of the field. All metrics were transformed to be collinear with log-odds-ratio between concordant and discordant calls. The model was refined using cross-validation. The logit function with the final logistic regression coefficients is being used to compute the scores in production.

LFR Haplotyping Algorithm.

Methods Figure 1 describes the main steps in the phasing of the LFR data. These steps are as follows:

1. Graph construction using LFR data: An undirected graph is made, where the vertices represent the heterozygous SNPs, and the edges represent the connection between those heterozygous SNPs. The edge is composed of the orientation and the strength of the connection.

2. Graph construction using mate pair: Similar to Step 1, where the connections are made based on the mate pair data, as opposed to the LFR data. For a connection to be made, a DNB must be found with the two heterozygous SNPs of interest in the same read (same arm or mate arm).

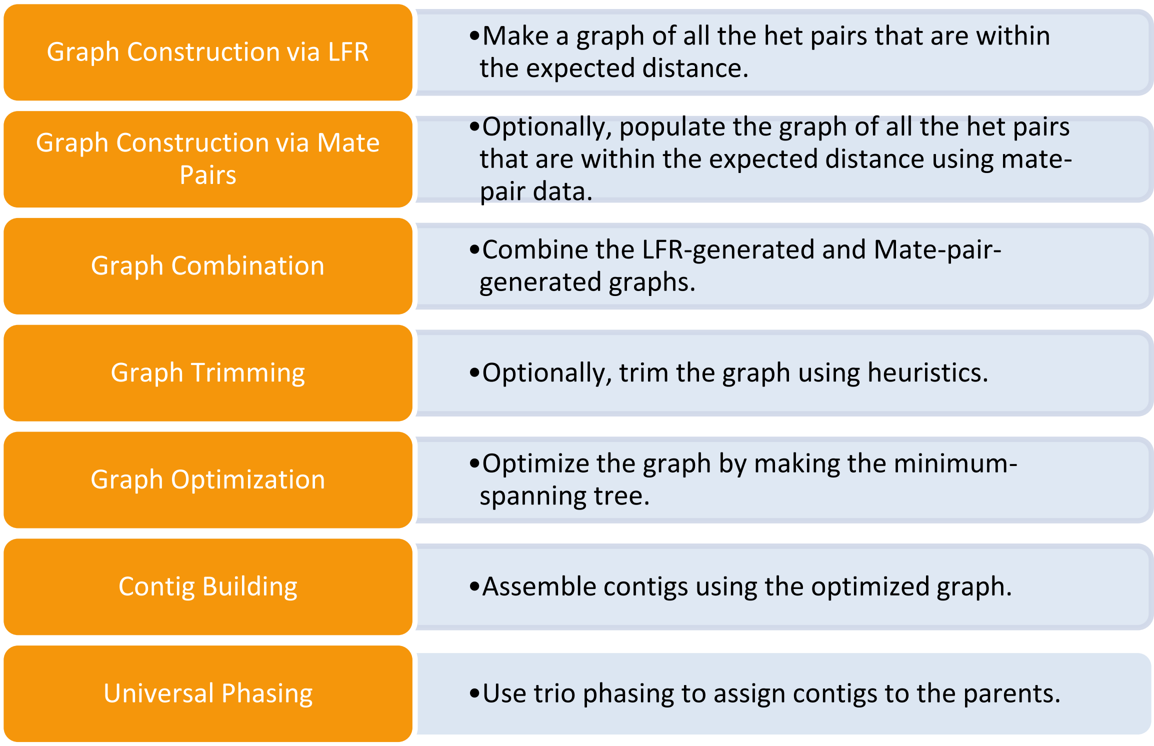
3. Graph combination: The representation of each of the above graphs is via an NxN sparse matrix, where N is the number of candidate heterozygous SNPs on that chromosome. Two nodes can only have one connection in each of the above methods. Where the two methods are combined, there may be up to two connections for two nodes. Therefore, a selection algorithm must be used to select one connection as the connection of choice. For these studies, we discovered that the quality of the mate-pair data was significantly inferior to that of the LFR. Therefore, only the LFR-derived connections were used.

4. Graph trimming: A series of heuristics were devised and applied in an attempt to remove some of the erroneous connections. More precisely, a node must satisfy the condition of at least 2 connections in one direction and 1 connection in the other direction; otherwise, it is eliminated.

5. Graph optimization: The graph was optimized by generating the minimum-spanning tree (MST). The energy function was set to -|strength|. During this process, where possible, the lower strength edges get eliminated, due to the competition with the stronger paths. Therefore, MST provides a natural selection for the strongest and most reliable connections.

6. Contig building: Once the minimum-spanning tree is available, all the nodes can be re-oriented with taking one node (here, the first node) constant. This first node is the anchor node. For each of the nodes, the path to the anchor node is found. The orientation of the test node is the aggregate of the orientations of the edges on the path.

7. Universal phasing: After the above steps, each of the contigs is phased. Here, we refer to the results of this part as pre-phased, as opposed to phased, indicating that this is not the final phasing. Since the first node was chosen arbitrarily as the anchor node, the phasing of the whole contig is not necessarily in-line with the parental chromosomes. For universal phasing, a few heterozygous SNPs on the contig for which trio information is available are used. These trio heterozygous SNPs are then used to identify the alignment of the contig. At the end of the universal phasing step, all the contigs have been labeled properly, and therefore can be considered as a chromosome-wide contig.

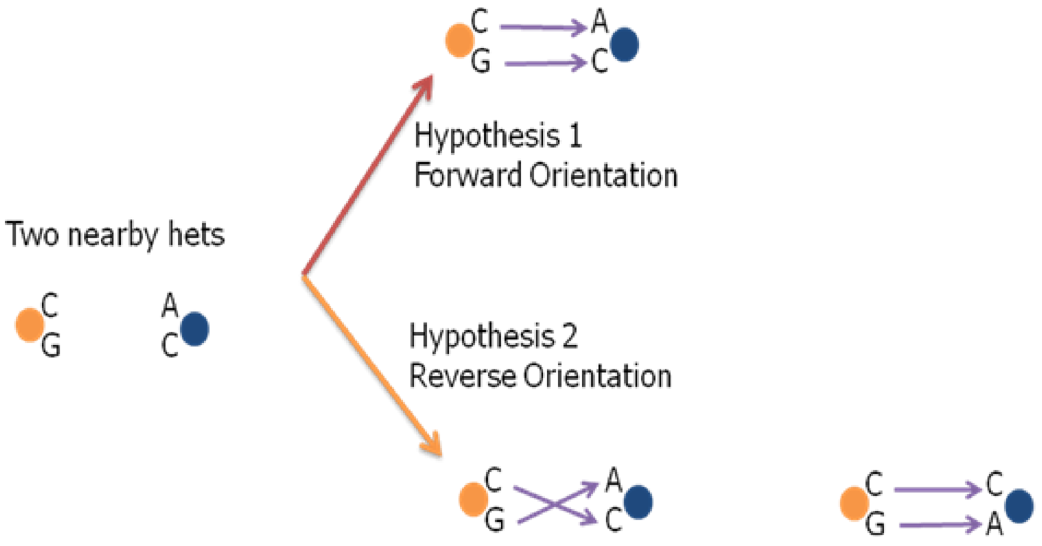

Graph Construction via LFR
•Make a graph of all the het pairs that are within the expected distance.
Graph Construction via Mate Pairs
•Optionally, populate the graph of all the het pairs that are within the expected distance using mate-pair data.
Graph Combination
•Combine the LFR-generated and Mate-pair-generated graphs.
Graph Trimming
•Optionally, trim the graph using heuristics.
Graph Optimization
•Optimize the graph by making the minimum-spanning tree.
Contig Building
•Assemble contigs using the optimized graph.
Universal Phasing
•Use trio phasing to assign contigs to the parents.

Methods Figure 1: General architecture of the LFR Algorithm

Contig Making

The crux of this algorithm is the 4x4 connectivity matrix. Each of the 4x4 cells of this connectivity matrix is populated with the support for that specific het combination. This support, in turn reflects the number of shared aliquots between the calls made for the two specific heterozygous calls corresponding to the connectivity matrix. For instance, suppose for Locus K, we are interrogating two hets --Het1 and Het2. Cell 1,2 (or Cell A/C) (reflecting an A for Het1, and C for Het2) is filled with the shared aliquots between the A call (number of wells that participated in making the A call) in Het1, and C call (number of wells that participated in making the C call) in Het2.

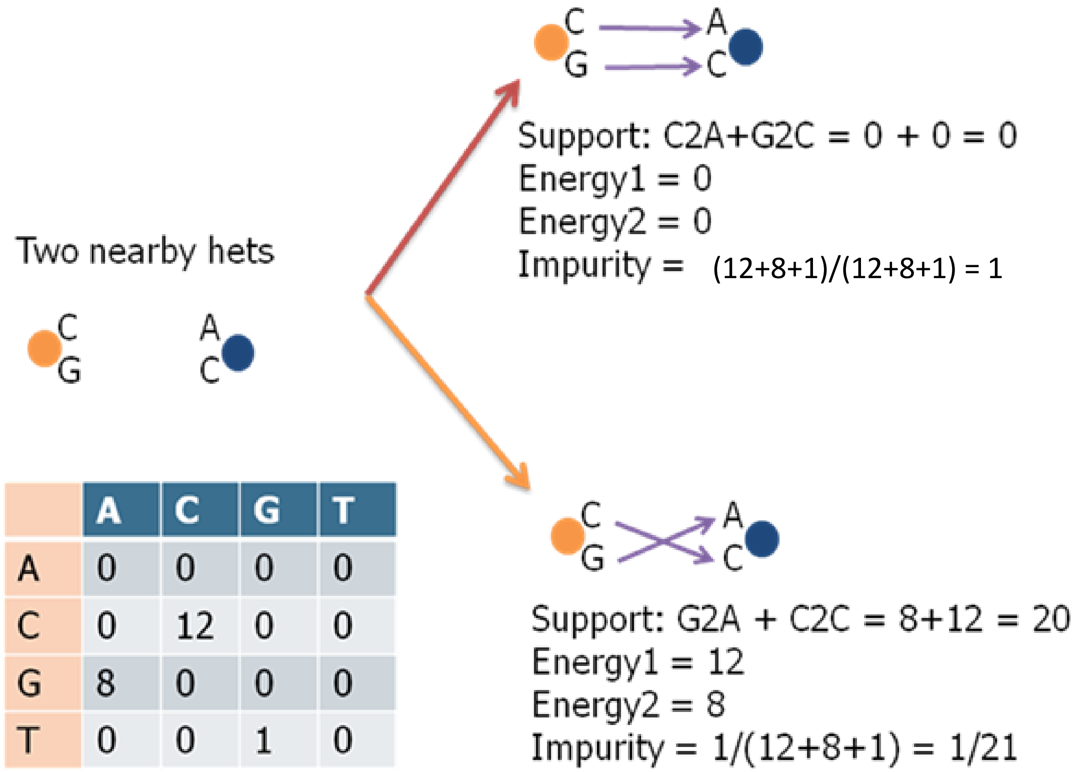
For each heterozygous SNP-pair, two hypotheses are tested --forward orientation and reverse orientation. A forward orientation means that the two heterozygous SNPs are connected the same way they are originally listed (initially alphabetically). A reverse orientation means that the two heterozygous SNPs are connected in reverse order of their original listing. Methods Figure 2 depicts the assignment of forward and reverse orientations to a heterozygous SNP-pair.



Methods Figure 2: Pairwise analysis of the nearby heterozygous SNPs

Each orientation will have a numerical support, showing the validity of the corresponding hypothesis. This support is a function of the 4x4=16 cells of the connectivity matrix shown in Methods Figure 3. To simplify the function, the 16 variables are reduced to 3 -- Energy1, Energy2 and Impurity. Energy 1 and Energy2 are two highest value cells corresponding to each hypothesis. Impurity is the ratio of the sum of all the other cells (than the two corresponding to the hypothesis) to the total sum of the cells in the matrix. In order to achieve longer contigs, up to 40% impurity was allowed in the current runs. Likely causes of impurity are: 1) wrong mapping due to the mapper error, 2) wrong mapping due to the genome compression (areas of the genome that have repeats, but appear as unique on the reference genome), 3) wrong tag decoding, 4) wrong raw basecalling (based on intensities), and 4) wrong het calling (false positives or false negatives).

The three metrics --Energy1, Energy2 and Impurity are fed into a fuzzy inference system (Methods Figure 4), in order to reduce their effects into a single value --score-- between (and including) 0 and 1.



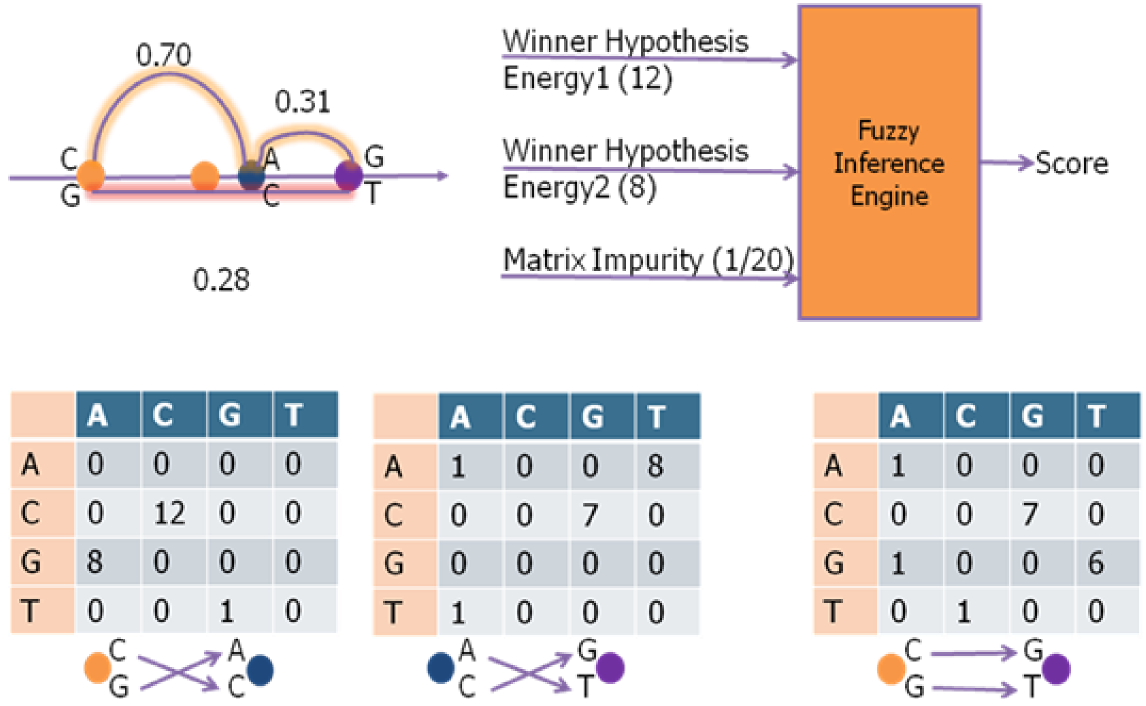
	A	C	G	T
A	0	0	0	0
C	0	12	0	0
G	8	0	0	0
T	0	0	1	0

Support: C2A+G2C = 0 + 0 = 0
Energy1 = 0
Energy2 = 0
Impurity = (12+8+1)/(12+8+1) = 1

Support: G2A + C2C = 8+12 = 20
Energy1 = 12
Energy2 = 8
Impurity = 1/(12+8+1) = 1/21

Methods Figure 3: An example of the selection of a hypothesis, and the assignment of a score to it.

The connectivity operation is done for each heterozygous SNP-pair that are within a reasonable distance up to the expected contig length (e.g., 20-50 Kb). Methods Figure 4 depicts some exemplary connectivities and strengths for 3 nearby heterozygous SNPs.



	A	C	G	T
A	0	0	0	0
C	0	12	0	0
G	8	0	0	0
T	0	0	1	0

	A	C	G	T
A	1	0	0	8
C	0	0	7	0
G	0	0	0	0
T	1	0	0	0

	A	C	G	T
A	1	0	0	0
C	0	0	7	0
G	1	0	0	6
T	0	1	0	0

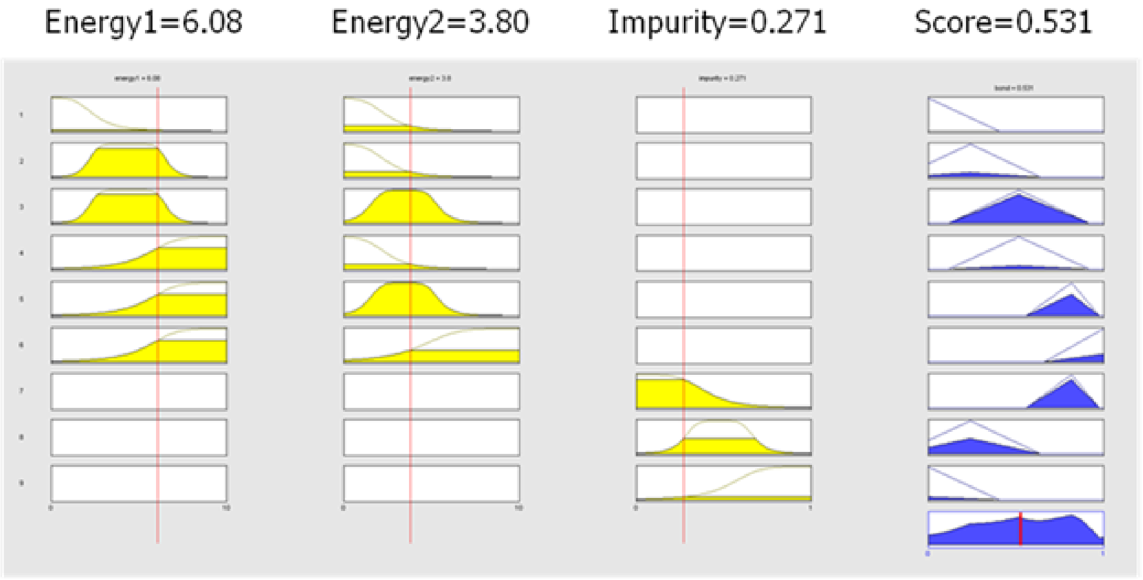
Methods Figure 4: Graph construction

The rules of the fuzzy inference engine are defined as follows:

1. If Energy1 is small and Energy2 is small, then Score is very small.
2. If Energy1 is medium and Energy2 is small, then Score is small.
3. If Energy1 is medium and Energy2 is medium, then Score is medium.
4. If Energy1 is large and Energy2 is small, then Score is medium.
5. If Energy1 is large and Energy2 is medium, then Score is large.
6. If Energy1 is large and Energy2 is large, then Score is very large.
7. If Impurity is small, then Score is large.
8. If Impurity is medium, then Score is small.
9. If Impurity is large, then Score is very small.

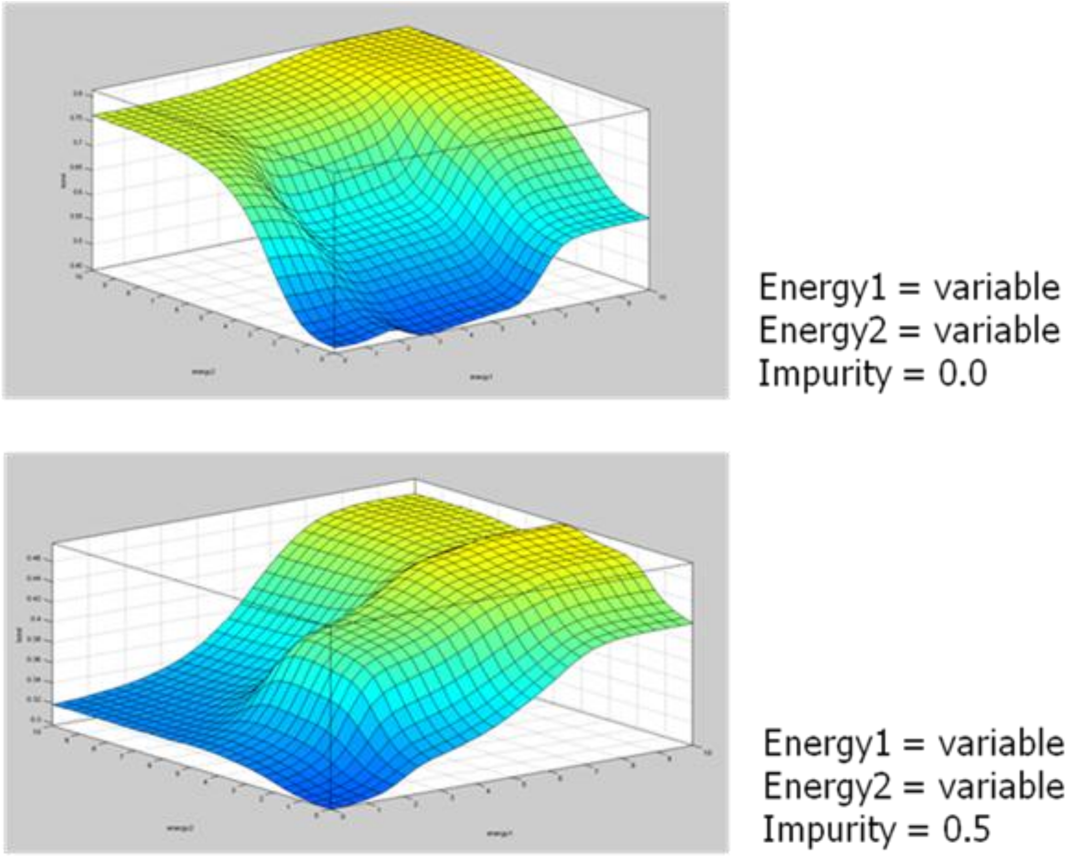
For each variable, the definition of Small, Medium and Large is different, and is governed by its specific membership functions. Methods Figure 5 shows the membership functions for each variable (in the vertical direction), along with an exemplary input/output scenario.

After exposing the FIS to each variable set, the contribution of the input set on the rules are propagated through the fuzzy logic system, and a single (de-fuzzified) number is generated at the output --score. This score is limited between 0 and 1, with 1 showing the highest quality. Methods Figure 5 demonstrates the incorporation of the the fuzzy rules, and the defuzzification process in the output, for one specific set of inputs.



Methods Figure 5: Fuzzy logic input/output example

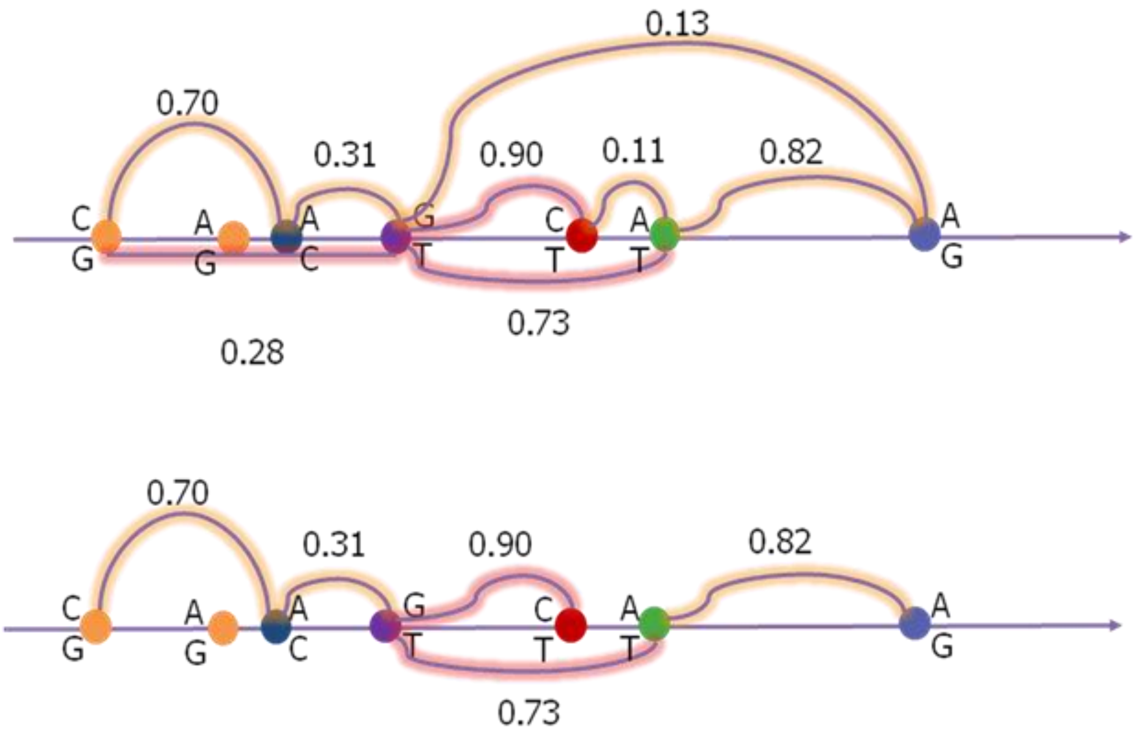
Methods Figure 6 shows the output curves for the fixed impurity values --0 and 0.5.



Methods Figure 6: Fuzzy logic input sweep

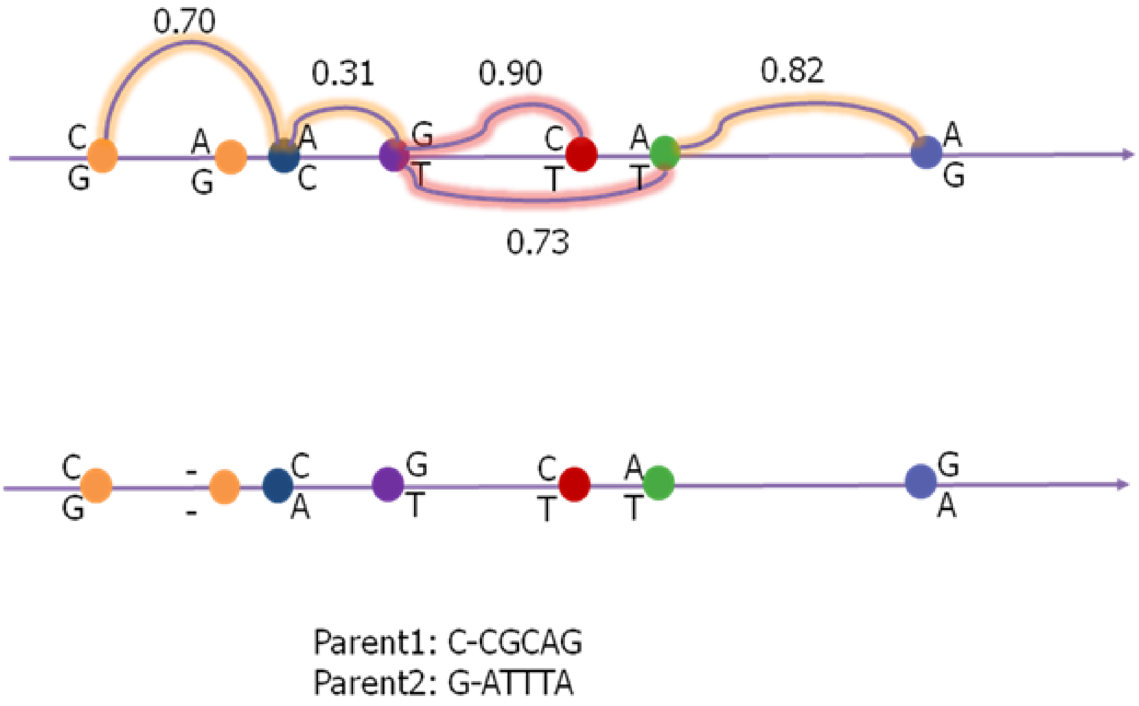
After the application of the FIS to each node pair, a complete graph is constructed.

Since hets are only expected to have fragment support while on the same fragment, het pairs were only considered for a distance given by the expected length of the fragments (here: 30 Kb). Also, for implementation reasons, het combinations were only considered for hets that were away from each other by a maximum number of hets (here 30). These two constraints help retaining the sparse nature for the all het-to-het matrix (of 4x4 connectivity matrices). Methods Figure 7 shows an example of such graph. The nodes are colored according to the orientation of the winning hypothesis. The strength of each connection is derived from the application of the FIS on the heterozygous SNP pair of interest. Once the preliminary graph is constructed (the top plot of Methods Figure 7), the graph is optimized (the bottom plot of Methods Figure 7) and reduced to a tree. This optimization process is done by making a Minimum Spanning Tree (MST) from the original graph. The MST guarantees a unique path from each node to any other node.



Methods Figure 7: Graph optimization

In this application, the first node on each contig is used as the anchor node, and all the other nodes are oriented to that node. Depending on the orientation, each hit would have to either flip or not, in order to match the orientation of the anchor node. Methods Figure 8 demonstrates such process for the given example. At the end of this process, a phased contig is made available.

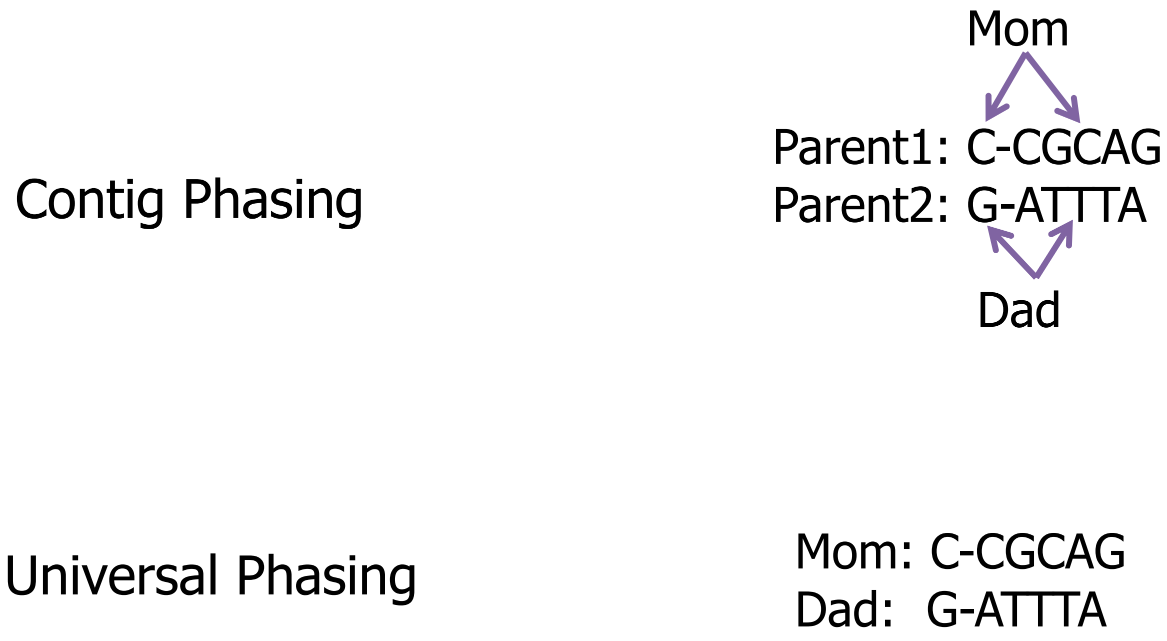


Methods Figure 8: Contig alignment

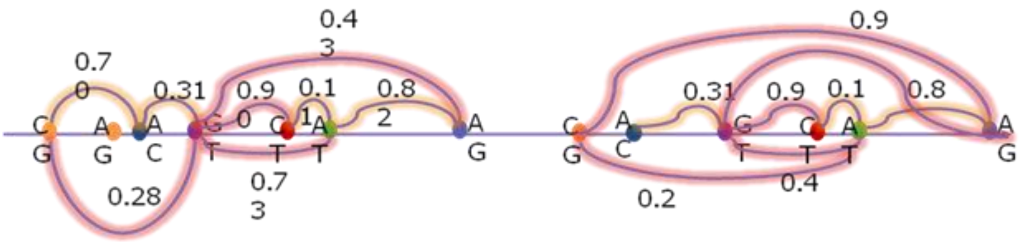
At this point, in the process of phasing, the two haplotypes are separated. Although it is known that one of these haplotypes comes from the Mom and one from the Dad, it is not known exactly which one comes from which parent. In the next step of phasing, we attempt to assign the correct parental label (Mom/Dad) to each haplotype. We refer to this process as the Universal Phasing. In order to do so, one needs to know the association of at least a few of the heterozygous SNPs (on the contig) to the parents. This information can be obtained by doing a Trio (Mom-Dad-Child) phasing. Using the trio's sequenced genomes, we identified some loci with known parental associations –more specifically when at least one parent is homozygous. These associations were then used to assign the correct parental label (Mom/Dad) to the whole contigs (Methods Figure 9).

In order to guarantee a high accuracy, we did the following: 1) when possible (e.g., in the case of NA19240), acquired the trio information from multiple sources (e.g., Internal and

1000Genomes), and used a combination of them; 2) required the contigs to have at least 2 known trio-phased loci on them; 3) eliminated the contigs that had a series of trio-mismatches in-a-row on them (indicating a segmental error); 4) eliminated the contigs that had a single trio-mismatch at the end of the trio loci (indicating a potential segmental error).



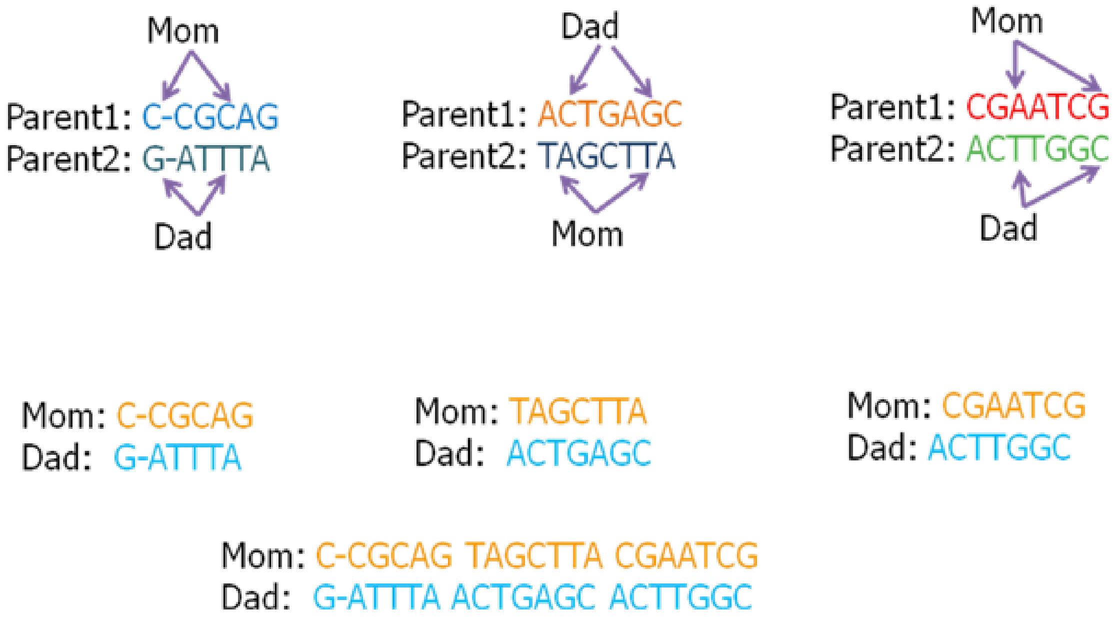
Methods Figure 9: Parent-assisted universal phasing



Methods Figure 10: Natural contig separations

Whether parental data is used or not, contigs often do not continue naturally beyond a certain point. Reasons for contig separation are: 1) More than usual DNA fragmentation or lack of amplification in certain areas, 2) Low heterozygous SNP density, 3) Poly-N sequence on the reference genome, and 4) DNA repeat regions (prone to mis-mapping).

One of the advantages of the Universal Phasing is the ability to obtain the full chromosomal "contigs." This is possible, since each contig (after the universal phasing) carries haplotypes with the correct parental labels. Therefore, all the contigs that carry the label Mom can be put on the same haplotype; and a similar operation can be done for Dad's contigs.



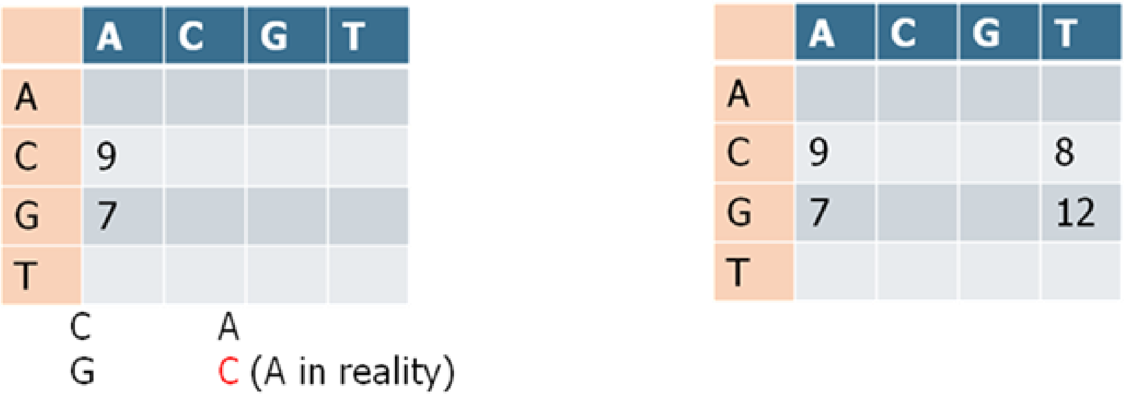
Methods Figure 11: Universal phasing

One of the major advantages of the LFR process is increasing the accuracy of the heterozygous SNP calling. Below, there are two examples of such correction:

1. Methods Figure 12a (left): The connectivity matrix does not support any of the expected hypotheses. This is an indication that one of the heterozygous SNPs is not really a heterozygous SNP. In this example, the A/C heterozygous SNP is in reality a homozygous locus (A/A), which was mislabeled as a heterozygous SNPerozygous locus by the assembler. This error can be identified, and either eliminated or (in this case) corrected.

2. Methods Figure 12b (right): The connectivity matrix for this case supports both hypotheses at the same time. This is a sign that the heterozygous SNPerozygous calls are not real.

A "healthy" heterozygous SNP-connection matrix is one that has only two high cells (at the expected heterozygous SNP positions, i.e., not on a straight line). All other possibilities point to potential problems, and can be either eliminated, or used to make alternate basecalls for the loci of interest.



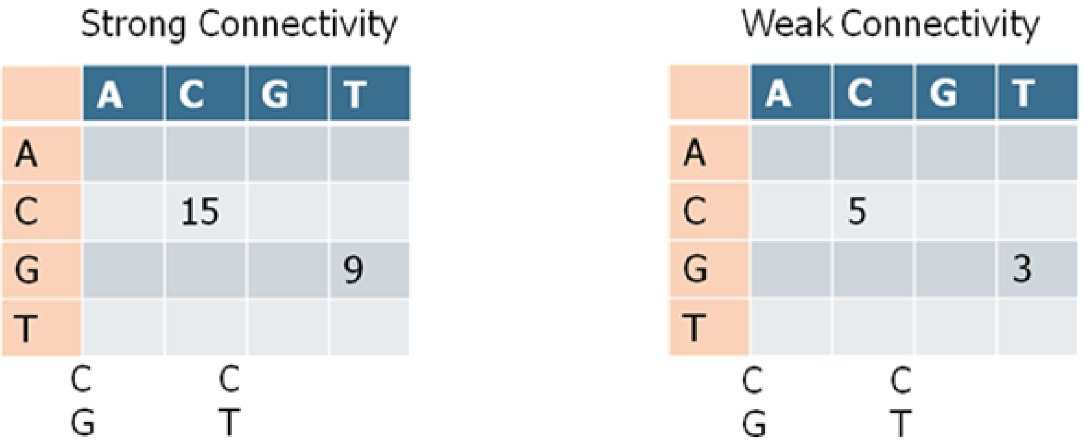
	A	C	G	T
A				
C	9			
G	7			
T				

C A
G C (A in reality)

	A	C	G	T
A				
C	9			8
G	7			12
T				

Methods Figure 12: Inherent error detection capability

Another advantage of the LFR process is the ability to call heterozygous SNPs with weak supports (e.g., where it was hard to map DNBs due to the bias or mismatch rate). Since the LFR process requires an extra constraint on the heterozygous SNPs, one could reduce the threshold that a heterozygous SNP call requires in a non-LFR assembler. Methods Figure 13 demonstrates an example of this case. In Methods Figure 13b (right) under a normal scenario the low number of supporting reads would have prevented any assembler to confidently call the corresponding heterozygous SNPs. However, since the connectivity matrix is "clean," one could more confidently assign heterozygous SNP calls to these loci.



Strong Connectivity

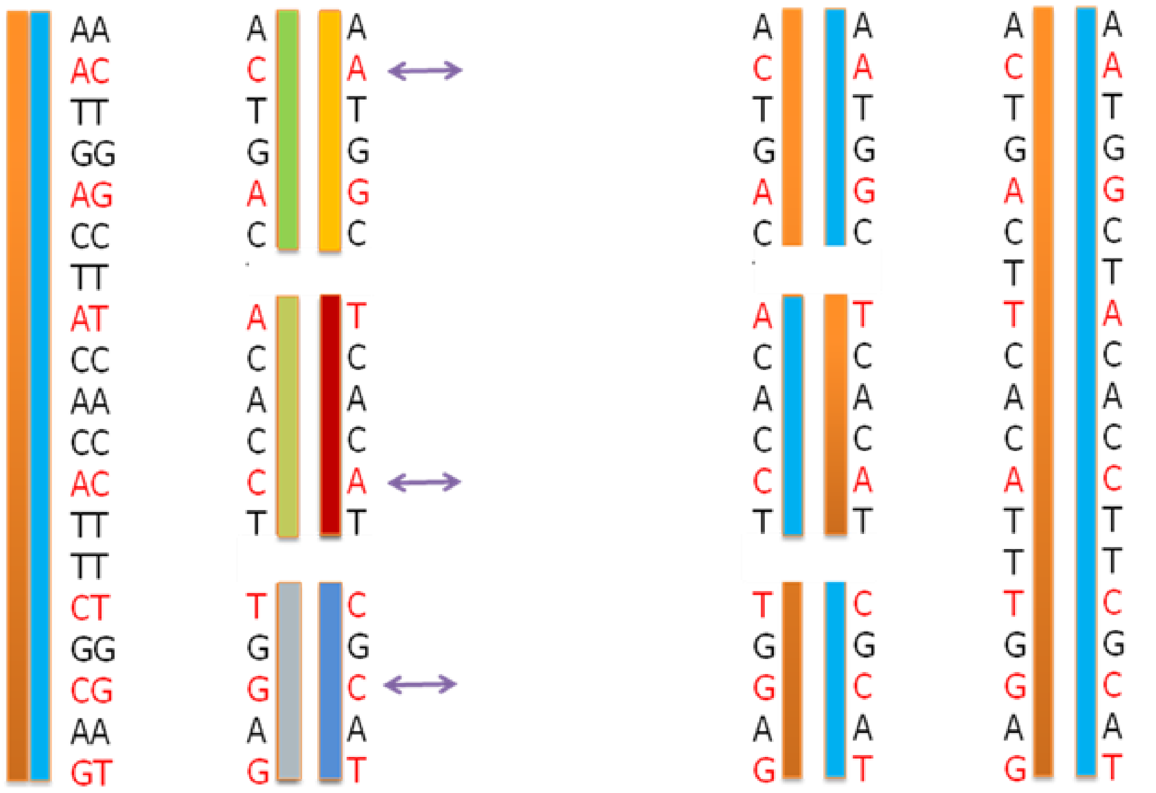
	A	C	G	T
A				
C		15		
G				9
T				

Weak Connectivity

	A	C	G	T
A				
C		5		
G				3
T				

Methods Figure 13: Decreasing false negatives. A confident heterozygous SNP call could be made despite a small number of reads.

Methods Figure 14 shows a summary of the LFR phasing process, including the step of universal phasing. Highlighted in red are the loci for which parental information has been available. On the right plots, the Mom and Dad chromosomes have been fully phased and shown with orange and blue colors, respectively.



Methods Figure 14: Phasing Process

Indel phasing process. To phase indels a 3x3 connectivity matrix with memberships for combinations of A/B/O alleles, where A is for Allele A, B is for B Allele, and O for Other Allele would be used. This representation supports both SNPs and indels. To enable this type of representation, the tally matrices should adopt the same data type, i.e., A, B and O should capture the counts of the reads for each well/aliquot at each locus for each of the A, B and O alleles, respectively. Beyond this alternate abstraction of the data, the rest will fit well into the existing pipeline, including 1) the fuzzy inference engine for evaluation of scores, 2) making of the graphs, 3) optimization of the graphs, and the remaining downstream processes/reports.

Annotating SNPs in Splice Sites. Introns in transcribed RNAs need to be spliced out before they become mRNA. Information for splicing is embedded within the sequence of these RNAs, and is consensus based. Mutations in splicing site consensus sequence are causes to many human diseases. It is well-established that the majority of splice sites conform to a simple consensus at fixed positions around an exon. We have developed a simple program to annotate Splice Site mutations. In this program, we used consensus splice position models from Steve Mount's database[33]. Simply, we are looking for a pattern: CAG|G in the 5'-end region of an exon ("|" denotes the beginning of exon), and MAG|GTRAG in the 3'-end region of the same exon ("|" denotes the ending of exon). Here M = {A,C}, R={A,G}. Further, splicing consensus positions are classified into 2 types. Type I, where consensus to the model is 100% required; and type II , where consensus to the model is preserved in >50% cases. Presumably, a SNP mutation in type I position will cause the splicing to miss, whereas a SNP in a type II position will only decrease the efficiency of the splicing event.

Our program is composed of two parts. In part I, we generate a file containing model positions sequences from the input reference genome. In part 2, we compare the SNPs from a sequencing project to these model sequences and report any type I and type II mutations. Our program is exon-centric instead of intron-centric (for the convenience in parsing the genome). For a given exon, in its 5'-end , we look for the consensus "cAGg " (for positions -3, -2, -1, 0. 0 means the start of exon). Capital letters means type I positions, and lower-case letters means type II positions). In the 3'-end of the exon, we look for the consensus "magGTrag" (for position sequence -3, -2, -1, 0, 1, 2, 3, 4). Exons from the genome release that do not confirm to these requirements are simply ignored

(~5% of all cases). These exons fall into other minor classes of splice-site consensus and are not investigated by this program. Any SNP from the genome sequenced is compared to the model sequence at these genomic positions. Any mismatch in type I will be reported. Mismatch in type II positions are reported if the mutation departs from the consensus.

It has to be noted that this program is a very simple program that detects the majority of bad Splice Site mutations. The bad SNPs reported are definitely problematic. But there are many other bad SNPs causing splicing problem are not detected by this program. For example, there are many introns within the human genome that do not confirm to the above-mentioned consensus. Also, mutations in bifurcation points in the middle of the intron may also cause splice problem. These Splice Site mutations are not reported.

Annotation of SNPs affecting Transcription Factor Binding Sites (TFBS). JASPAR models are used for finding TFBSs from the released human genome sequences (either build 36 or build 37). JASPAR Core is a collection of 130 TFBS positional frequency data for vertebrates, modeled as matrices[34,35]. These models were downloaded from JASPAR website. These models were converted into Position Weight Matrices (PWMs) using the following formula:

$$wi = \mathrm{log2}\ [(fi+p\ Ni1/2)\ /(Ni+\ Ni1/2)/p]$$

where fi is the observed frequency for the specific base at position i, and Ni is the total observations at the position; and p the background frequency for the current nucleotide, which is defaulted to 0.25. A specific program, mast[36], is used to search sequence segments within the genome for TFBS-sites.

We first run a program to extract TFBS-sites in the reference genome. Here is the outline of steps:

i) For each gene with mRNA, extract [-5000, 1000] putative TFBS-containing regions from the genome, with 0 being the mRNA starting location.

ii) Run mast-search of all PWM-models for the putative TFBS-containing sequences

iii) Select those hits above a given threshold.

iv) For regions with multiple or overlapping hits, we select only 1-hit, the one with the highest mast-search score.

With the TFBS model-hits from the reference genome in hand, we can now identify SNPs which are located within the hit-region. These SNPs will impact on the model, and a change in the hit-score. A second program was written to compute such changes in the hit-score, as the segment containing the SNP was run twice into the PWM model, once for the reference, and the second time for the one with the SNP substitution. A SNP causing the segment hit score to drop more than 3 is identified as a detrimental SNP.

Selection of genes with 2 detrimental SNPs. Genes with detrimental SNPs are classified into 2 categories: 1) Those affecting the AA-sequence transcribed. 2) Those affecting the transcription binding site. For AA-sequence affecting, we include the following SNP subcategories:

1) NONSENSE or NONSTOP variations. These mutations either cause a truncated protein or an extended protein. In either situation, the function of the protein product is either completely lost or less efficient.

2) Splice Site variations. These mutations cause either the splice site for an intron to be destroyed (for those positions required to be 100% of a certain nucleotide by the model) or severely diminished (for those sites required to be >50% for a certain nucleotide by the model. The SNP causes the splice-site nucleotide to mutate to another nucleotide that is below 50% of consensus as predicted by the splice-site consensus sequence model). These mutations will likely produce proteins which are truncated, missing exons, or severely diminishing in protein product quantity.

3) Polyphen2[31] annotation of AA variations. For SNPs that cause change in amino-acid sequence of a protein, but not its length, we use Polyphen2 as our main annotation tool. Polyphen2 annotates the SNP with "benign", "unknown, "possibly damaging", and "probably damaging". We identify both "possibly damaging" and "probably damaging" as bad SNPs. These category assignments by Polyphen2 are based on structural predictions of the Polyphen2 software.

For transcription-binding site mutations, we used the 75% of maxScore of the models based on the reference genome as a screening for TFBS-binding sites. Any model-hit in the region that is <=75% of maxScore are removed. For those remaining, if a SNP causes the hit-score to drop 3 or more, we take it as a detrimental SNP.

We report 2 classes of genes. Class 1 genes are those that had at least 2-bad AA-affecting mutations. These mutations can be all on a single allele (Class 1.1), or spread on 2 distinct alleles (Class 1.2).

Class 2 genes are a superset of the Class 1 set. Class 2 genes are genes contain at least 2 detrimental SNPs, irrespective it is AA-affecting or TFBS-site affecting. But we require at least 1 SNP is AA-affecting. Class 2 genes are those either in Class 1, or those that have 1 detrimental AA-mutation and 1 or more detrimental TFBS-affecting variations. Class 2.1 means that all these detrimental mutations are from a single allele, whereas Class 2.2 means that detrimental SNPs are coming from 2 distinct alleles.