# Phasing the Chromosomes of a Family Group When One Parent is Missing

T. Whit Athey

## Abstract

A technique is presented for the phasing of sets of SNP data collected from members of a family group consisting of at least three siblings and at least one parent. The process works best if data for four or more siblings is available. When data for one parent is missing, the phased chromosomes for that parent are reconstructed as a part of the process. Examples are presented from a family group consisting of one parent and four children, and the recombination diagrams for three example chromosomes are included. Having the recombination diagrams for the family group allows the identification of the grandparent of the siblings through whom any given DNA segment has passed.

### Introduction

Genetic testing by means of a "gene chip" has become popular as a way to assess approximate relative susceptibility to various diseases and also as a tool in genealogical research. For each of the 22 pairs of autosomal chromosomes, the nucleotide or base at a half-million to a million locations is reported by the testing procedure. The chips are single use and are discarded following use. There are two main manufacturers of the chips, Illumina and Affymetrix (Wikipedia, 2010).

The locations of the reported bases along the chromosomes, or single nucleotide polymorphisms (SNPs), are chosen for inclusion on the gene chip because of their variability in the human population. At a random location in the human genome, humans are greater than 99% identical, but at the locations of the SNP set on the typical gene chip, humans are about 75% identical.

For each SNP the testing process returns two results since the chromosomes exist in pairs with one chromosome of the pair coming from each parent. In general, at a given SNP location it is usually impossible to determine which of the two bases reported for a particular SNP is on which chromosome. The attempt to separate the two sets of bases into their natural grouping according to which chromosome they reside on is termed the "phasing" of the chromosomes.

Address for correspondence: T. Whit Athey, wathey@hprg.com

Received: June 22, 2010; Accepted: Nov. 14, 2010; Published December 19, 2010



Once the chromosomes of a family group are phased, the drawing of recombination diagrams becomes easy to carry out. The recombination diagram for a particular chromosome shows where the DNA of each child came from. It is also possible to use matches with cousins to identify the origin of each of the parental chromosomes—exactly which grandparent each child's DNA came from. This is very useful in tracing genealogical relationships to distantly related persons as will be explained in more detail below.

The technique presented here is intended to address the problem of phasing the chromosomes where three or more siblings, preferably four or more, plus one parent have available SNP raw data from one of the companies offering gene chip testing services such as 23andMe, Navigenics, DeCodeMe, or Family Tree DNA. The suggested approach also provides the reconstructed and phased chromosomes of the missing parent, at least to the extent that the missing parent has actually passed along every part of both of his chromosomes to at least one of the children. If this latter condition is not strictly met, the phasing and reconstruction can still be carried out, but the missing part of the chromosome will still be missing from the reconstructed data.

On average a parent will not pass ¼ his/her DNA in the case of two siblings, 1/8 in the case of three siblings, 1/16 in the case of four siblings, etc. However, this average value reflects many chromosomes where no DNA has not been passed to any sibling and others where more than the average has not been passed. In the regions where DNA from one of the chromosomes of the missing parent has not been passed down, the reconstructed chromosome will have a block of missing data, though the other reconstructed chromosome will be available.

**Methods**

For convenience we first number the phasing principles that can be applied at a single location—the "point location logic" for phasing. The first two of these are well known and form the basis of "classical" phasing approaches.

Principle 1 -- If a person is "homozygous" at a location—that is, having the same base on each of the two chromosomes of a pair, then obviously at that location it is possible to know with certainty that both chromosomes of the pair have that base at that location, but this is an almost trivial form of phasing.

Principle 2 -- If data from one of the parents are available, and that parent is homozygous at a SNP location, then another almost trivial phasing is possible since obviously that parent had to send the only type of base s/he had at that location to the child.

Principle 3 -- A final phasing principle is almost trivial, but it is normally not useful because there is usually no way to satisfy its conditions: If a child is heterozygous at a particular SNP, and if it is possible to determine which parent contributed one of the bases, then the other parent necessarily contributed the other (or alternate) base. This principle will be very useful in the present approach.

However, there is an additional source of information for the phasing problem, at least when data for several siblings are available, and this principle represents the main new insight into the phasing problem to be presented here. We know that there will only be a small number of cuts and crossovers during the process of recombination. Recombination is the process where a parent cuts and pastes together his/her two chromosomes of a pair to produce one hybrid chromosome to pass along to the offspring. In general, the sequence on the single chromosome passed from one parent will remain from one chromosomal origin (of that parent's two chromosomes) for millions or tens of millions of bases before a crossover point is reached. So, if one can deduce the source (chromosome) of a base (i.e., which parental chromosome) in a sequence or in a series of individual SNP locations, the neighboring bases or SNP results in that set will usually be from the same parental chromosome—the principle of inheritance of DNA in "haploblocks." Combining the principles of "point location logic" and "haploblock inheritance" has made it possible to develop a technique that allows the phasing of genomic data from a family group where data from three (ideally four) or more siblings is available, along with data from at least one parent.

The technique presented here involves an iterative procedure wherein successive passes through the data from the participants gradually produces more and more of the bases that each child received from their two parents (including the bases of the missing parent). The process is best explained by use of an example. In this example, the missing parent is the father, so the mother's data is included at the beginning and the father's data is reconstructed in the process.

Included in the following table are just a few dozen rows taken from a much larger spreadsheet that contained all the reported SNP data for chromosome 16 for each of the five family members. The small number of rows shown here (out of the 17000+ that were available) was limited for ease of presentation. The two columns for each family member are just labeled "left" and "right" at this point since we don't know their origin. The raw data are shown in Table 1.

On the basis of the values in Table 1, we can construct a crude partial phasing at some of the SNP locations, namely those locations where one of the siblings has the same base on both chromosomes (homozygous), or where the tested parent has the same base on both chromosomes. These locations will not initially be very informative, but can provide a start toward complete phasing. For example, for the above data, at a few locations, we can show the base that the father must have contributed to each of the children and the base that the mother must have contributed to each of the children. Many locations will not contain useful information or will contain no information at all and are left blank. After completing this first step we can show which base was contributed to each of the four children by the parents for a few locations as shown in Table 2.

Note the pattern of inheritance from Dad shown in Table 2 for the four siblings in the leftmost four columns. The first few rows show an AABB base pattern, but this gives way in about lines 12-13 to a new pattern, ABBB. Even though we only can see the pattern showing in some of the rows, these patterns persist over hundreds or thousands of SNPs, and can be assumed to exist also in the intervening rows where no pattern was discernable (and in the underlying sequence). Note that often there will be the same base in every location, a case of "accidental matching" which does not contribute to or detract from the pattern we are looking for. When two or more bases are different in a row, however, this represents an informative pattern—if any two are different, then since there are only two possible chromosomes contributing, it means we can see the chromosomal origins of the bases.

The above example was chosen to include the transition from one pattern to another. We will look at the

| SNP | Location | Mom Left | Mom Right | Sib1 Left | Sib1 Right | Sib2 Left | Sib2 Right | Sib3 Left | Sib3 Right | Sib4 Left | Sib4 Right |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SNP1 | 4931055 | C | C | C | C | C | C | C | C | C | C |
| SNP2 | 4939393 | A | A | A | A | A | A | A | G | A | G |
| SNP3 | 4941305 | C | T | C | T | T | T | T | T | C | T |
| SNP4 | 4941381 | C | C | C | C | C | C | C | C | C | C |
| SNP5 | 4950914 | G | T | G | G | G | T | G | T | G | G |
| SNP6 | 4957388 | T | T | C | T | C | T | C | T | C | T |
| SNP7 | 4961984 | T | T | C | T | C | T | T | T | T | T |
| SNP8 | 4963062 | G | G | G | G | G | G | A | G | A | G |
| SNP9 | 4963227 | G | G | A | G | A | G | G | G | G | G |
| SNP10 | 4963271 | C | C | C | C | C | C | A | C | A | C |
| SNP11 | 4963522 | C | C | A | C | A | C | C | C | C | C |
| SNP12 | 4966485 | T | T | T | T | T | T | T | T | T | T |
| SNP13 | 4966592 | C | T | C | T | C | C | C | C | C | T |
| SNP14 | 4967765 | G | G | A | G | G | G | G | G | G | G |
| SNP15 | 4971845 | A | A | A | A | A | A | A | A | A | A |
| SNP16 | 4973983 | C | C | C | T | C | C | C | C | C | C |
| SNP17 | 4974880 | A | A | A | G | A | G | A | G | A | G |
| SNP18 | 4979249 | C | C | C | C | C | C | C | C | C | C |
| SNP19 | 4989315 | G | T | G | T | G | T | G | T | G | G |
| SNP20 | 4993332 | T | T | T | T | T | T | T | T | T | T |
| SNP21 | 4996296 | T | T | C | T | T | T | T | T | T | T |
| SNP22 | 5000569 | C | T | C | T | C | T | C | T | C | C |
| SNP23 | 5001972 | C | T | C | T | C | T | C | T | C | C |
| SNP24 | 5004229 | C | C | C | C | C | C | C | C | C | C |
| SNP25 | 5004918 | G | G | G | G | G | G | G | G | G | G |
| SNP26 | 5008153 | A | G | A | G | A | A | A | A | A | G |
| SNP27 | 5008751 | A | G | A | G | A | A | A | A | A | G |
| SNP28 | 5009356 | A | A | A | A | A | A | A | A | A | A |
| SNP29 | 5011671 | A | G | A | G | A | A | A | A | A | G |
| SNP30 | 5013774 | C | T | C | T | C | C | C | C | C | T |
| SNP31 | 5014953 | C | C | C | C | C | C | C | C | C | C |
| SNP32 | 5015543 | C | C | A | C | A | C | A | C | A | C |
| SNP33 | 5017453 | T | T | C | T | C | T | C | T | C | T |
| SNP34 | 5022232 | T | T | C | T | T | T | T | T | T | T |
| SNP35 | 5030996 | G | G | G | G | G | G | G | G | G | G |
| SNP36 | 5032119 | T | T | C | T | C | T | C | T | C | T |
| SNP37 | 5035600 | C | C | C | T | C | C | C | C | C | C |
| SNP38 | 5045441 | G | G | A | G | A | G | A | G | A | G |
| SNP39 | 5050272 | C | C | C | T | C | C | C | C | C | C |
| SNP40 | 5050627 | C | C | A | C | C | C | C | C | C | C |
| SNP41 | 5055229 | G | G | G | T | G | G | G | G | G | G |
| SNP42 | 5059873 | T | T | C | T | T | T | T | T | T | T |
| SNP43 | 5062769 | G | G | A | G | G | G | G | G | G | G |

Table 1: Raw Data From Part of Chromosome 16 for the Family Group

mother's pattern in the next step, but first we will fill in as much as possible of the father's contribution on the basis of the data we have already, and on the basis of the father's pattern that we have identified that characterizes each region. For example, consider the third row of Table 2, which contains for Dad's contribution, the

| SNP Numbers | Sib1 From Dad | Sib2 From Dad | Sib3 From Dad | Sib4 From Dad | Dad Informative Pattern | Sib1 From Mom | Sib2 From Mom | Sib3 From Mom | Sib4 From Mom |
|---|---|---|---|---|---|---|---|---|---|
| SNP1 | C | C | C | C | | C | C | C | C |
| SNP2 | A | A | G | G | AAGG | A | A | A | A |
| SNP3 | | T | T | | | | T | T | |
| SNP4 | C | C | C | C | | C | C | C | C |
| SNP5 | G | | | G | | G | | | G |
| SNP6 | C | C | C | C | | T | T | T | T |
| SNP7 | C | C | T | T | CCTT | T | T | T | T |
| SNP8 | G | G | A | A | GGAA | G | G | G | G |
| SNP9 | A | A | G | G | AAGG | G | G | G | G |
| SNP10 | C | C | A | A | CCAA | C | C | C | C |
| SNP11 | A | A | C | C | AACC | C | C | C | C |
| SNP12 | T | T | T | T | | T | T | T | T |
| SNP13 | | C | C | | | | C | C | |
| SNP14 | A | G | G | G | AGGG | G | G | G | G |
| SNP15 | A | A | A | A | | A | A | A | A |
| SNP16 | T | C | C | C | TCCC | C | C | C | C |
| SNP17 | G | G | G | G | | A | A | A | A |
| SNP18 | C | C | C | C | | C | C | C | C |
| SNP19 | | | | G | | | | | G |
| SNP20 | T | T | T | T | | T | T | T | T |
| SNP21 | C | T | T | T | CTTT | T | T | T | T |
| SNP22 | | | | C | | | | | C |
| SNP23 | | | | C | | | | | C |
| SNP24 | C | C | C | C | | C | C | C | C |
| SNP25 | G | G | G | G | | G | G | G | G |
| SNP26 | | A | A | | | | A | A | |
| SNP27 | | A | A | | | | A | A | |
| SNP28 | A | A | A | A | | A | A | A | A |
| SNP29 | | A | A | | | | A | A | |
| SNP30 | | C | C | | | | C | C | |
| SNP31 | C | C | C | C | | C | C | C | C |
| SNP32 | A | A | A | A | | C | C | C | C |
| SNP33 | C | C | C | C | | T | T | T | T |
| SNP34 | C | T | T | T | CTTT | T | T | T | T |
| SNP35 | G | G | G | G | | G | G | G | G |
| SNP36 | C | C | C | C | | T | T | T | T |
| SNP37 | T | C | C | C | TCCC | C | C | C | C |
| SNP38 | A | A | A | A | | G | G | G | G |
| SNP39 | T | C | C | C | TCCC | C | C | C | C |
| SNP40 | A | C | C | C | ACCC | C | C | C | C |
| SNP41 | T | G | G | G | TGGG | G | G | G | G |
| SNP42 | C | T | T | T | CTTT | T | T | T | T |
| SNP43 | A | G | G | G | AGGG | G | G | G | G |

Table 2: Initial Partial Phasing of the Raw Data for the Four Siblings

following bases (two of which are unknown):

? T T ?

However, we know that the first two bases must be the same from the pattern, and also that the second two bases must be the same from the pattern. Therefore, the four

bases become completely filled in as:

T T T T

Similarly, in the 19th row of Table 2, which is mostly empty, and which is past the point where the pattern changes to ABBB, we have only:

? ? ? C

However, because the pattern tells us that the rightmost three bases must all be equal (because they are coming from the same parental chromosome), then the four bases become:

? C C C

which is considerably more filled in than it was initially. Also in each row, we can recall that the empty (blank) cells resulted only from cases where the bases were different in what the parents possessed, or in what the child possessed. In completing as much as we can in each row, we leave any filled cells alone. For some of the unfilled cells on the mother's side of the table, we can fill in the alternative (other) base from the corresponding location on the father's side of the table. That is, we know that the sibling with an empty cell got one base from the father, but the alternative base from the mother. Therefore, after the use of the Dad pattern fills in more cells, a newly filled-in cell in the father's side of the table gives rise to a filled-in cell in the same position on the mother's side--the alternative base to what was on the father's side. In the table below, the shading in the leftmost four columns reminds us of the patterns and where they change. After applying those two operations discussed above to each row in the table, the new version of the table is shown in Table 3.

Note that we could only fill in cells on the left on the basis of those positions that are equal. That is, if the Dad's pattern is ABBB, and if we have the cells partially filled in as:

? ? G G

then we can extend this on the basis of the pattern to:

? G G G

but we cannot extend it to T G G G. This is because the first base remains ambiguous—it could be a G, since it could match by "accident," coming from the other chromosome. The second base must be a G, however, because those positions that are equal in the pattern (because they come from the same chromosome) cannot be other than equal, or else a new pattern is established. That is, bases can match "by accident," but they cannot be different "by accident"—differences definitely mean different chromosomes. Since we do not have direct access to the father's data, we must leave this location uncertain and unfilled for now. Note that this restriction would not apply in applying the pattern to the mother's side of the table because we do have her data.

Note that the cells in Table 3 are now considerably more filled compared to Table 2.

In the next step, we use the pattern on the mother's side to fill in as many more cells as possible. Finally, we can project the information in those newly filled cells back to the father's side using Principle 3 again. In this example, the same pattern for the mother exists throughout the whole table, but this would not be true for a table representing a whole chromosome in general.

In the rightmost column in the above table, the pattern for the Mom's contribution is shown, though there are only two rows with informative data concerning the pattern. Adopting this pattern would be problematic if based upon just two rows, but this example shows only a few rows of the larger spreadsheet, whereas in that larger spreadsheet the mother's pattern could be followed over many hundreds of rows.

There will usually be some uncertainty as to the exact location where a pattern changes. Often the approximate points where one should be on the lookout for a pattern change may be found from the locations of the half-identical segments shown in the Excel comparison tables that can be downloaded from the testing company. These half-identical segments, based on genotypes, tend to run somewhat longer than segments based on haplotypes.

After we fill in the mother's contribution columns as much as possible from the mother's pattern, we then use those newly filled-in cells to fill in any corresponding empty cell on the father's side with the alternative or complementary base. After those processes, the table becomes as shown in Table 4.

In filling in the cells in Table 4 in this last step, note that the right (mother's) side was filled in first, followed by filling in any possible additional cells on the left. The two cells that are left unshaded in column 2 on the left represent cells where the base is known, but it is unknown exactly where the crossover point from one pattern to another may be. The location of the crossover will be often be ambiguous within a few SNP locations.

Note that our procedure has completely filled in the table in this example. The contributions of each parent to each child is explicitly shown. In practice, about 1-2% of the cells may remain unfilled. This occurs when there is missing data for the mother, or where every family member is heterozygous in exactly the same way. This small amount of missing information will not usually affect the phasing of the overall chromosomes. The

| SNP Numbers | Sib1 From Dad | Sib2 From Dad | Sib3 From Dad | Sib4 From Dad | Dad Informative Pattern | Sib1 From Mom | Sib2 From Mom | Sib3 From Mom | Sib4 From Mom | Mom's Informative Pattern |
|---|---|---|---|---|---|---|---|---|---|---|
| SNP1 | C | C | C | C | | C | C | C | C | |
| SNP2 | A | A | G | G | AAGG | A | A | A | A | |
| SNP3 | T | T | T | T | | C | T | T | C | CTTC |
| SNP4 | C | C | C | C | | C | C | C | C | |
| SNP5 | G | G | G | G | | G | T | T | G | GTTG |
| SNP6 | C | C | C | C | | T | T | T | T | |
| SNP7 | C | C | T | T | CCTT | T | T | T | T | |
| SNP8 | G | G | A | A | GGAA | G | G | G | G | |
| SNP9 | A | A | G | G | AAGG | G | G | G | G | |
| SNP10 | C | C | A | A | CCAA | C | C | C | C | |
| SNP11 | A | A | C | C | AACC | C | C | C | C | |
| SNP12 | T | T | T | T | | T | T | T | T | |
| SNP13 | | C | C | C | | | C | C | T | |
| SNP14 | A | G | G | G | AGGG | G | G | G | G | |
| SNP15 | A | A | A | A | | A | A | A | A | |
| SNP16 | T | C | C | C | TCCC | C | C | C | C | |
| SNP17 | G | G | G | G | | A | A | A | A | |
| SNP18 | C | C | C | C | | C | C | C | C | |
| SNP19 | | G | G | G | | | T | T | G | |
| SNP20 | T | T | T | T | | T | T | T | T | |
| SNP21 | C | T | T | T | CTTT | T | T | T | T | |
| SNP22 | | C | C | C | | | T | T | C | |
| SNP23 | | C | C | C | | | T | T | C | |
| SNP24 | C | C | C | C | | C | C | C | C | |
| SNP25 | G | G | G | G | | G | G | G | G | |
| SNP26 | | A | A | A | | | A | A | G | |
| SNP27 | | A | A | A | | | A | A | G | |
| SNP28 | A | A | A | A | | A | A | A | A | |
| SNP29 | | A | A | A | | | A | A | G | |
| SNP30 | | C | C | C | | | C | C | T | |
| SNP31 | C | C | C | C | | C | C | C | C | |
| SNP32 | A | A | A | A | | C | C | C | C | |
| SNP33 | C | C | C | C | | T | T | T | T | |
| SNP34 | C | T | T | T | CTTT | T | T | T | T | |
| SNP35 | G | G | G | G | | G | G | G | G | |
| SNP36 | C | C | C | C | | T | T | T | T | |
| SNP37 | T | C | C | C | TCCC | C | C | C | C | |
| SNP38 | A | A | A | A | | G | G | G | G | |
| SNP39 | T | C | C | C | TCCC | C | C | C | C | |
| SNP40 | A | C | C | C | ACCC | C | C | C | C | |
| SNP41 | T | G | G | G | TGGG | G | G | G | G | |
| SNP42 | C | T | T | T | CTTT | T | T | T | T | |
| SNP43 | A | G | G | G | AGGG | G | G | G | G | |

Table 3: Filling in More of the Father's Contribution

columns have been shaded with color to show the origin of the chromosomes for each child. On the mother's side of the table, the same pattern existed throughout this example region, so the first and fourth child got the mother's "green" chromosome, while the second and third child got the mother's "dark blue" chromosome.

| SNP Numbers | Sib1 From Dad | Sib2 From Dad | Sib3 From Dad | Sib4 From Dad | Sib1 From Mom | Sib2 From Mom | Sib3 From Mom | Sib4 From Mom |
|---|---|---|---|---|---|---|---|---|
| SNP1 | C | C | C | C | C | C | C | C |
| SNP2 | A | A | G | G | A | A | A | A |
| SNP3 | T | T | T | T | C | T | T | C |
| SNP4 | C | C | C | C | C | C | C | C |
| SNP5 | G | G | G | G | G | T | T | G |
| SNP6 | C | C | C | C | T | T | T | T |
| SNP7 | C | C | T | T | T | T | T | T |
| SNP8 | G | G | A | A | G | G | G | G |
| SNP9 | A | A | G | G | G | G | G | G |
| SNP10 | C | C | A | A | C | C | C | C |
| SNP11 | A | A | C | C | C | C | C | C |
| SNP12 | T | T | T | T | T | T | T | T |
| SNP13 | C | C | C | C | T | C | C | T |
| SNP14 | A | G | G | G | G | G | G | G |
| SNP15 | A | A | A | A | A | A | A | A |
| SNP16 | T | C | C | C | C | C | C | C |
| SNP17 | G | G | G | G | A | A | A | A |
| SNP18 | C | C | C | C | C | C | C | C |
| SNP19 | T | G | G | G | G | T | T | G |
| SNP20 | T | T | T | T | T | T | T | T |
| SNP21 | C | T | T | T | T | T | T | T |
| SNP22 | T | C | C | C | C | T | T | C |
| SNP23 | T | C | C | C | C | T | T | C |
| SNP24 | C | C | C | C | C | C | C | C |
| SNP25 | G | G | G | G | G | G | G | G |
| SNP26 | A | A | A | A | G | A | A | G |
| SNP27 | A | A | A | A | G | A | A | G |
| SNP28 | A | A | A | A | A | A | A | A |
| SNP29 | A | A | A | A | G | A | A | G |
| SNP30 | C | C | C | C | T | C | C | T |
| SNP31 | C | C | C | C | C | C | C | C |
| SNP32 | A | A | A | A | C | C | C | C |
| SNP33 | C | C | C | C | T | T | T | T |
| SNP34 | C | T | T | T | T | T | T | T |
| SNP35 | G | G | G | G | G | G | G | G |
| SNP36 | C | C | C | C | T | T | T | T |
| SNP37 | T | C | C | C | C | C | C | C |
| SNP38 | A | A | A | A | G | G | G | G |
| SNP39 | T | C | C | C | C | C | C | C |
| SNP40 | A | C | C | C | C | C | C | C |
| SNP41 | T | G | G | G | G | G | G | G |
| SNP42 | C | T | T | T | T | T | T | T |
| SNP43 | A | G | G | G | G | G | G | G |

Table 4: Phased Chromosomes for the Siblings and Parents

On the father's side, rows 1-12 show that the first two children got the "light blue" chromosome from their father, while the third and fourth children got the "red" chromosome from their father. After the 13th row where the father's pattern has changed, the second child's chromosome switches to the father's red chromosome, while the chromosomes of the first, third, and fourth child are continuous (in grandparent origin) across the boundary of the two patterns.

At this point the chromosomes are completely phased for the entire family group, including the absent father. The father's chromosomes have been reconstructed as a part of this process.

It is sometimes helpful to repeat the last two steps to fill in a few more missing values.

Sometimes, even with four or more children, some (usually small) parts of the missing parent's chromosomes may not be passed to any of the children. Although we can fill in the table as above and see what each child got from each parent, if some part of a chromosome was not passed to any child, it obviously can not be reconstructed. The other chromosome can be reconstructed, however. If the data from fewer than four siblings are used, it becomes more likely that part of one or both of the chromosomes of the missing parent will not have been passed to any of the children. A more serious problem with fewer sibling participants is that the inheritance patterns become less easy to detect and it is more difficult to follow a particular chromosome, for example one that has been identified as paternal or maternal, through many different changes in the pattern. With only two siblings, a quarter of the missing parents' chromosome on average will not have been passed to any offspring and the inheritance pattern will have only same or different possibilities and it becomes impossible to determine which sibling's chromosome is changing at a recombination cut point.

Although the chromosomes for each parent have been phased or separated, we can only apply arbitrary labels to them in general—we don't know whether the "light blue" chromosome represents the father's paternally derived chromosome or his maternally derived chromosome. However, if data for a cousin with a known relationship to this family group is available, it will often be possible to demonstrate which is the paternal and maternal chromosomes of the parent.

**Results and Discussion**

With the phased chromosomes in hand, it is a simple matter to construct recombination diagrams for the family group. In fact, the columns of Table 4 (if separated slightly) can be imagined as representing a schematic diagram of each chromosome of the four children. Simply re-grouping the columns into pairs by child, whether retaining the base names or not, provides a "diagram" of the recombination that has occurred in the family. Typically, chromosome diagrams are shown in a horizontal orientation, rather than vertical as in the columns of Table 4, and the columns for a real chromosome would be much too large to display as a whole on a page, but the chromosomes can easily be redrawn in a graphing or drawing software package for ease of display. Figures 1-3 show the recombination diagrams for our example family group for chromosomes 1, 6, and 15, using all of the available data (not just the abbreviated example used above). Such diagrams contain a wealth of information.

In these figures the data of a double-second cousin was used to identify each chromosome from the parents as being the paternally derived or maternally derived chromosome. This cousin's mother was independently a first cousin to each of the parents in the example family group. As a result of the known genealogy, where this cousin shared a segment with the mother of this group, the segment had to be on the mother's paternal chromosome. Where this cousin shared a segment with one or more of the siblings, but not the mother, then the segment had to be on the father's paternal chromosome. The two remaining parental chromosomes could then be labeled as maternal. On a few chromosomes, such as the three illustrated in Figures 1-3, both types of match occurred on the same chromosome, and this fact is highlighted near the bottom of each figure. Further testing of two additional second cousins is expected to resolve the identity (paternal/maternal) of most of the remaining chromosomes.

In the example chromosome 16 analyzed in this article, it was possible not only to show how each parent recombined their two chromosomes in passing one along to each child, but also it was possible to show which grandparent that any segment came from. This can be quite useful when a match is found to a new person in a company database where the relationship is unknown. It may be helpful to compare his raw data (if available) to the phased chromosomes and establish the pathway of descent through one of the four grandparents. This narrows the focus for trying to discover the relationship. Often this determination can be deduced by simply comparing the half-identical segment information and the recombination diagram, rather than comparing the raw data (see the bottom of Figure 2, for example).

If a shared segment is found with a cousin, whether or not the genealogical relationship with the main family

group is known, it will often be possible to use the raw data from such a person to fill in a few more of the missing phased values. One would first check the other person's raw data against the phased chromosomes of the mother and father in the family group. In the region where the segment is shared, the chromosome that is "related" can be identified with a half-identity calculator. Within that segment, we know that it is extremely likely that all of the segment was passed from one chromosome, so if there are any missing values in the parents' phased chromosomes, the missing values can sometimes be filled in from the cousin's data.

Once the recombination diagrams have been constructed, they apply equally well to any other SNP set from the same people, or even to their complete sequences. That is, the recombination diagrams are properties of the genomes being analyzed and are not specific to a particular SNP set.

The process described here has been implemented in an Excel spreadsheet, but a web-based program is planned.

**Limitations**

There are two practical limitations and one special case to the use of the approach described here.

1. Wherever the recombination process in one parent has used a closely matching cut point in two different children, it may be difficult to trace the four chromosomes through that point of changing inheritance patterns. It is important that only one sibling's chromosome changes across a cut point in order to trace all the chromosomes across that point. This is especially important when data from only three siblings is used—one must assume that at a point where a pattern changes, the chromosome of only one sibling has switched its source while the chromosomes of the other two siblings is continuous across that point.

2. When the pattern changes from a situation where three siblings have the same parental chromosome source and one sibling is different, to the situation where all four have the same source, it is difficult to decide where the pattern has made this change. There will be many SNPs where all four siblings have the same base, even when not all have come from the same parental chromosome. As a practical matter, it will be expedient to just take the last SNP with an informative pattern as the end of that pattern and the beginning of the all-same pattern. However, the resulting effect on the other parent's table in the beginning of the all-same region will need to be examined to see if it forces any discrepant or rapidly changing patterns.

3. The X chromosome presents a special case for phasing, unless all of the siblings are female. Males have only one X chromosome that they received from their mothers, so that one X does not need phasing. However, it will still be of interest to determine the recombination that the mother carried out. The X chromosomes of any female siblings will need to be phased as described here. Matches with cousins can still show the origin of the X chromosomes of the mother, but the X that the father had must necessarily be from the father's mother. The approach presented here will need to be modified to take account of the specific male-female makeup of the family group, and it will be important that the mother's data be available.

**References**

Wikipedia (2010) Gene chip. URL: http://en.wikipedia.org/wiki/Gene_chip

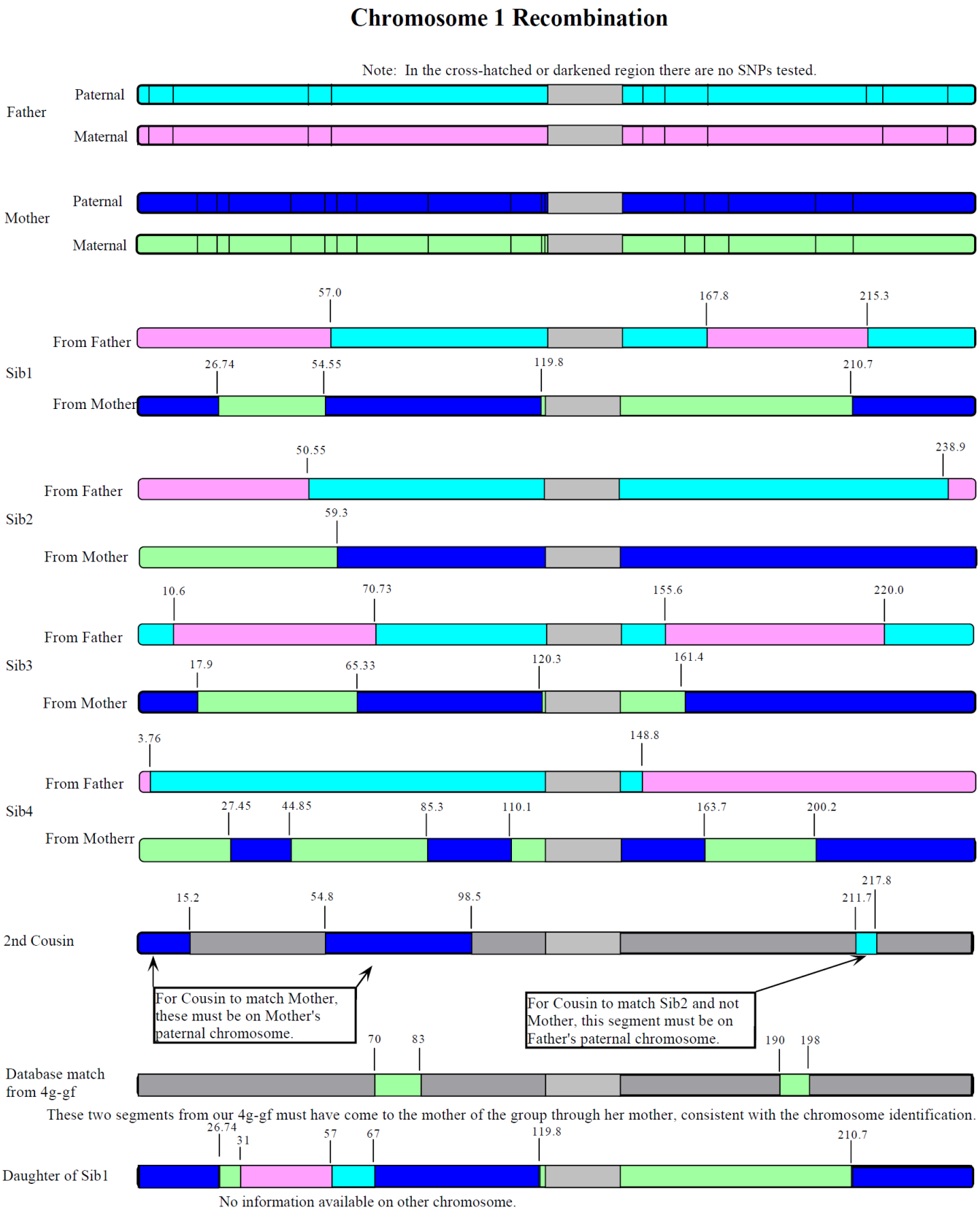


Figure 1: Recombination Diagram for the Family Group for Chromosome 1

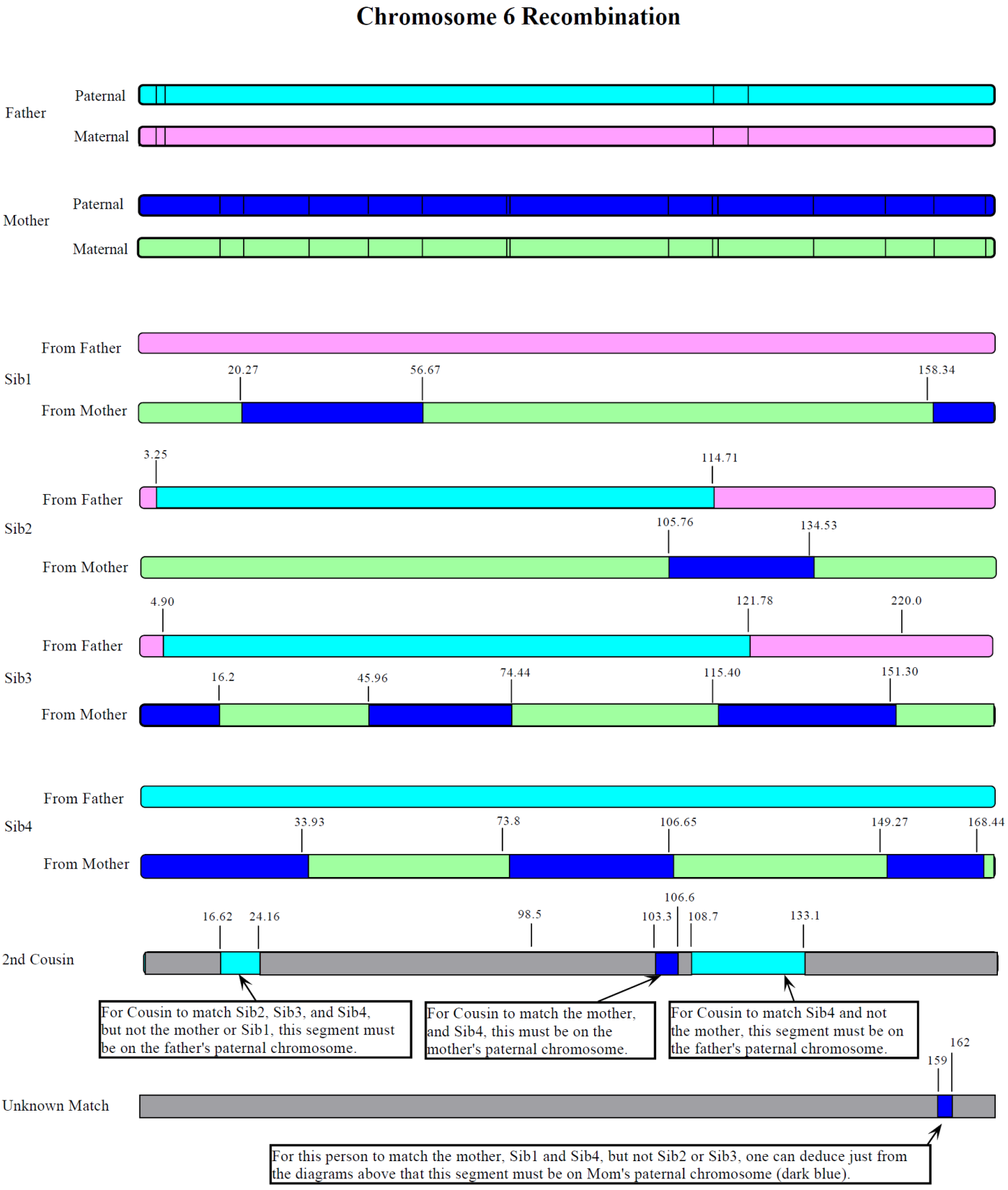


Figure 2: Recombination Diagram for Family Group for Chromosome 6

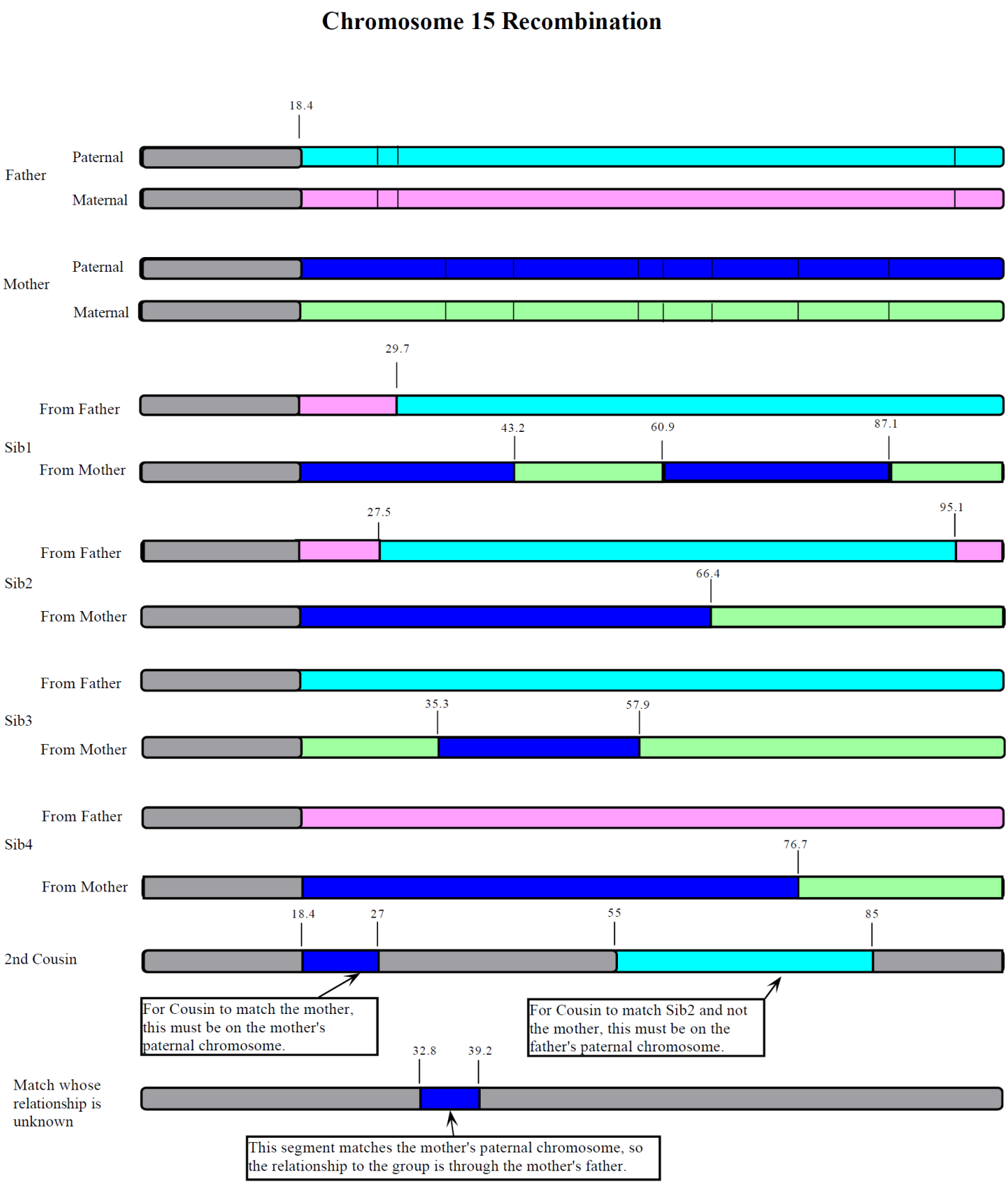


Figure 3: Recombination Diagram for Family Group for Chromosome 15