Mapping the Human Y Chromosome

by

Charles A. Tilford

B.A., Chemistry
Williams College (1991)

Submitted to the Department of Biology in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy in Biology

at the

Massachusetts Institute of Technology

June 2001

Signature of Author _____
Department of Biology
27 March 2001

Certified by _____
David C. Page, M.D.
Professor of Biology
Thesis Supervisor

Accepted by _____
Terry Orr-Weaver, Ph.D.
Professor of Biology
Chairperson, Biology Graduate Committee

Table of Contents

Mapping the Human Y Chromosome

by

Charles A. Tilford

Submitted to the Department of Biology on 9 May 2001
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Biology

Abstract

In humans, as well as in many other organisms, the Y chromosome stands out from the rest of the genome. Half the population lacks it entirely, yet it is essential for the proper development of the remaining half. Bearing the testis determining factor, inheritance of the Y chromosome is the first step in the path towards male development. Later in life, other factors on the chromosome are required for successful spermatogenesis. Further, the Y has proven invaluable to population geneticists and evolutionary biologists, as much of it is recombinationally isolated from the rest of the genome.

In spite of its unique biology and genetics, genomic studies on the Y chromosome have always been beset by more difficulties than the rest of the genome. The non-recombining portion can not be linkage mapped, and is also exceptionally rich in Y-specific repetitive sequences, making marker development challenging. In addition, many researchers dismiss the Y chromosome as a gene-poor curiosity, and focus their studies elsewhere in the genome.

We report here the construction of high-resolution radiation hybrid (RH) and physical maps of the Y chromosome. Initial efforts focused on developing new sequence-tagged sites (STSs) to provide adequate marker density along the chromosome. These STSs were developed by a novel clone-based subtraction protocol designed to eliminate both genome-typical and Y-specific repetitive sequences. Characterization of the markers indicates that the subtraction functioned as desired, such that the vast majority of the markers generated are single copy or present in distinct local clusters. The one limitation of the subtraction is its reliance on YAC clones, which are frequently chimeric.

These low-copy markers were then typed on the TNG RH panel, along with previously identified markers and BAC end markers developed during map construction. Assembly of the RH map required knowledge of prior mapping data. This was necessitated by the need to grossly order markers

manually before attempting to order the markers algorithmically. Because of the inclusion of repetitive markers, a naive global analysis of the data would result in inappropriate linkage between distant regions that shared repetitive markers. Geographically limited analysis of the data avoided this problem and allowed local order of markers to be reliably determined.

In the process of map construction we developed a novel algorithm to allow linkage analysis between markers of unspecified copy number. It was hoped that this software would allow more accurate determination of linkage between high-copy markers. The algorithm performs well, as compared to actual Y sequence, but does not provide a significant improvement over standard haploid models of linkage likelihood.

The RH map was used to select locally grouped markers for pooled screening of BAC filters. Clones identified in these screens were then verified and ordered in a physical map of the Y chromosome, which was in turn used to select BACs for sequencing. Sequencing of the entire non-recombining portion of the Y was completed in early 2001, excluding massively repetitive regions and four small gaps. This achievement represents the ultimate map of the chromosome.

Thesis Supervisor: David C. Page

Title: Professor of Biology

Member, Whitehead Institute for Biomedical Research

Investigator, Howard Hughes Medical Institute

Chapter 1: Introduction

Y Chromosome Biology and Structure

The Y chromosome is unique within the human genome. When comparing euchromatin, it is the smallest of the chromosomes, while preliminary analysis of other chromosomes indicates that it also contains the smallest number of gene families. Though gene duplication is occasionally observed on other chromosomes (Lupski, 1998; Mazzarella & Schlessinger, 1997), it is the norm on the Y. While all other chromosomes can pair with and engage in recombination along the length of their pairing partner, the Y will only pair and recombine with the X at the very distal tips. It is the only chromosome that may be completely absent and still result in the development of a viable adult. On the other hand, Y ploidy in excess of 3 chromosomes is also consistent with viability (Shanske *et al.*, 1998), a characteristic shared only by the X chromosome.

These features can be largely understood as a product of the evolutionary history that generated the Y chromosome. Experimental evidence suggests that roughly 300 million years ago the X and Y chromosomes functioned as an autosomal pair, and were thus indistinguishable(Lahn & Page, 1999a). Sex determination in this ancestral organism would not have been chromosomal, and would instead have relied on alternate methods for partitioning gender among the progeny. At some point one member of the pair, the future Y, acquired a dominant testis determining factor (TDF). This function, recognized today in the gene SRY, likely arose through mutation of a pre-existing gene in the sex determination pathway, given that while the initiation of sex determination varies between

species, the molecular components are often conserved (Western *et al.*, 2000). SOX-3 on the X chromosome shows some homology to SRY and is likely the survivor from the ancestral, non-mutated gene.

After the chromosome (now the future Y) acquired TDF, it underwent a series of inversions (Lahn & Page, 1999a). Each event suppressed recombination within the inverted area, which in turn led to the gradual degradation of sequences within the region. As with any alteration in the genome, selection determined what mutations would be allowed through evolutionary time, and which would be selected against. Unlike recombining chromosomes, however, mutations within the non-recombining portion of the Y chromosome (the NRY) are presented to the species as a "take it or leave it" option. If a deleterious mutation occurs in the context of an otherwise advantageous autosome, it may be selectively removed from the gene pool by recombination. Within the NRY, a harmful mutation occurring in a genomic context that also contains beneficial variations is then irreversibly linked to the "good" genes in that area (Rice, 1987). The chromosome will only be selected against if the *net* benefit of the chromosome is less than other chromosomes in the population.

If the overall impact is instead detrimental, then the species can respond only by compensating for the deleterious mutation elsewhere in the genome (aside from eliminating carriers of the chromosome by selection). In the case of the evolving Y, a frequently expected mutation will be the total disruption of the Y copy in an XY gene pair, such that functional mRNA is produced only from the X chromosome. This will typically result in dosage differences between the genders, with males producing half the level of transcript as females. This dosage problem is common to many different species with chromosomally-based sex determination. In mammals, it has been solved by

X inactivation, whereby cells with more than one X chromosome undergo a chromosome-wide inactivation of all but one of the X chromosomes. This assures that females have only one transcriptionaly active X, and thus produce mRNA levels comparable to XY males (Charlesworth, 1996).

Occasionally, elimination of a Y gene is so deleterious that these mutations are immediately removed from the population. There are several XY gene pairs that survive within the non-recombining region of the chromosome (Lahn & Page, 1997). They all appear to be housekeeping genes, such that a sudden drop to half levels of mRNA transcription would presumably be extremely deleterious or lethal. Since two copies of the gene survive in males, and transcription level is critical, females would likewise require expression of two copies. While the mechanism is unknown, it is observed that in all cases these surviving XY pairs "escape from X inactivation" in females - while the majority of the inactive X undergoes global transcriptional silencing, those genes that have Y homologues are either re-activated or manage to avoid the initial inactivation. This balance between inactivation of whole portions of the X, while protecting discrete genes within that area, would likely have evolved gradually in concert with the degradation of the Y (Charlesworth, 1998; Jegalian & Page, 1998).

From these observations, the non-recombining portion of the Y chromosome should continue to be whittled down until it consists only of TDF and those few housekeeping XY-homologous genes that are highly dosage sensitive (and one might imagine that even these might be gradually eased into oblivion). Inspection of the chromosome has revealed a second class of inhabitants, however. These appear to be relatively new immigrants to the sex chromosomes, since their homologues are found not on the X, but rather on autosomes. Given that movement of discrete genetic elements can

occur within the genome through transposition, it is not surprising that the Y will have some such arrivals. What is striking is that these newcomers appear to impart a selective advantage, in that in many cases they not only are maintained over evolutionary time, but are also frequently amplified once present on the Y (Lahn & Page, 1999b; Saxena *et al.*, 1996). Such capture by the Y chromosome of testis-specific genes followed by amplification has also been observed in *Drosophila* (Kalmykova *et al.*, 1997).

Many of these Y-specific repeat families are still in the early stages of analysis, but strong support is emerging that as a group they function in spermatogenesis. All are expressed in the testis, sometimes exclusively (Lahn & Page, 1997). The most extensively studied, the DAZ gene family, is found to be wholly or partially deleted in 13% of men tested with complete spermatogenic failure (Reijo *et al.*, 1995), and when mutated in model organisms results in gametogenic failure (Eberhart *et al.*, 1996; Karashima *et al.*, 2000).

From a practical point of view, it might be most efficient for mammals to collect spermatogenesis genes on the Y chromosome. As females have no use for such genes, this simplifies regulation and provides some metabolic advantage in the form of less genetic material to maintain. From the standpoint of a genome with 6Gb of DNA and over 30,000 genes, however, such an advantage would likely be minor. It is more appealing to argue that, rather than providing a small efficiency benefit when stored only in males, these genes exert a large negative effect when present in females. This sexually antagonistic hypothesis would presume that while expression of these genes benefits males, they lead to a loss of fitness in females.

There is anecdotal evidence to support this theory. Occasionally, a 46,XY embryo fails to develop as a male, and instead proceeds with female

development. These 46,XY females are in all characteristics truly female, although they tend to be taller than their XX counterparts, in some cases there is elevated production of androgens, and in nearly all cases the patient is sterile (Mittwoch, 1992; Scully, 1970). The later is thought to be an effect of lacking the second X chromosome, as Turner patients (45,X0) show the same phenotype (Ogata & Matsuo, 1995). Like Turner patients, the vast majority of germ cells are lost in 46,XY females soon after birth. In a Turner individual, this depletion is complete, and the ovaries develop as a small, germ cell-free body of tissue referred to as a streak gonad (Singh & Carr, 1965). However, in the 46,XY female isolated germ cells survive and eventually proliferate. In at least 25-40% of such women this growth is uncontrolled (Schellhas, 1974), and in the absence of surgical intervention results in gonadoblastoma, a germ cell tumor virtually unheard of in the normal population (Schellhas, 1974; Scully, 1970; Scully, 1953).

Clearly then, the Y chromosome bears one or more factors that allow the ultimate survival of the germ cells, and likely also induces those cells to become cancerous. In several cases 46,XY gonadoblastoma patients have been identified with only a partial Y chromosome; combining such observations indicates that the factor(s) required for gonadoblastoma progression are likely located on the proximal short arm, and exclude SRY (Salo *et al.*, 1995; Tsuchiya *et al.*, 1995). A likely candidate within this region is TSPY, which is present in 20-40 copies in humans, and is expressed within gonadoblastoma cells (Tsuchiya *et al.*, 1995). TSPY shows minor homology to SET, an autosomal protein which has been associated with fusion proteins of HRX in some forms of acute leukemia (Adler *et al.*, 1997). However, other genes are present within the gonadoblastoma critical region, and to-date have not been ruled out. Regardless of its identity, whichever protein is

responsible for the cancerous transition of germ cells in 46,XY females may play a role in the normal proliferation of germ cells in males. This hypothesis seems reasonable given the dramatic difference in gametogenesis between the sexes - while females start reproductive life with a fixed number of gametes, mature males maintain a rapidly dividing population of stem cells that generate an extraordinary 130 million gametes each day (Sharpe, 1994). The factors responsible for this difference must be tightly controlled, as improper expression of male proliferative factors in the ovary could potentially result in cancer. Sequestration of such factors on the Y chromosome would assure that they would be absolutely absent in females, and would allow the proliferative advantage to be maximally realized in males. This argument also provides a consistent explanation for gonadoblastoma in 46,XY females, and its rarity in the general population.

If many of the Y-specific genes are then in fact spermatogenesis factors, their duplication on the chromosome may also be more easily explained. There is no consensus on what constitutes a "normal" sperm count, but quoted values typically range from 20-70 million sperm per ml of ejaculate (Silber, 2000). While this figure may seem a fantastic overexuberance on the part of reproductive evolution, values less than 20 million per ml are considered low enough to be potentially causative for infertility, although the correlation is contested by some researchers. Accepting this relationship implies that, within limits, fertility will increase as a function of the level of spermatogenesis. Since fertility is the ultimate selective factor in any species (if procreation is not possible, all other selective advantages are moot) most any mutation that increases the rate of spermatogenesis should be strongly selected for. The increase in transcription of rate-limiting spermatogenic factors should then be favored. In the previous discussion of dosage

compensation it was noted that alteration of copy number generally has a corresponding impact on the level of transcription, so it would then follow that duplication of Y chromosome spermatogenic factors may impart a selective advantage to that male.

Given that the actual function and mechanism of operation of the majority of Y genes is still unknown, this argument is largely speculative. An alternative hypothesis is that multiple copies of critical Y genes provides a mechanism for maintaining the fidelity of these genes in the absence of recombination. This would most likely involve gene conversion or a similar mechanism that would allow one copy to be used as a template to correct another, and would almost certainly be non-directed (that is, a defective copy could also overwrite a functional one), and therefore ultimately rely on selection to eliminate uncorrected deleterious mutations. It is also possible that the observed duplications serve no function whatsoever, and are simply a consequence of the absence of recombination. Finally, it is certainly possible that several, or none, of the above theories could be operating on the Y. In any case, gene duplication is not limited to the human Y chromosome - TSPY amplification has been seen in other mammals (up to 200 copies in *Bos taurus* (Jakubiczka *et al.*, 1993)), and amplification of DAZ exons within Old World monkeys have been observed as well (Gromoll *et al.*, 1999).

It has already been mentioned that deletion of all or portions of the DAZ gene cluster can lead to infertility in human males. Such gross rearrangement of duplicated genes falls into a class of genetic diseases that have been termed "genomic disorders" to reflect the rather extreme modification of the DNA (Lupski, 1998). These disorders arise when gene conversion or inappropriate recombination occur between repetitive elements in the genome. Typically, such an event requires near-identical sequence

conservation between the two regions involved, and is more likely as the length of the segments increases. The Y is extraordinarily rich with large, highly conserved sequences, so it is highly likely that deletion of DAZ will be but one of multiple genetic disorders ultimately described on the chromosome. Examples of disease-associated genomic rearrangements elsewhere in the genome include red-green color blindness (recombination within a tandem array), steroid-sulfatase deficiency (recombination between distant repetitive elements), and hemophilia A (recombination between inverted repeats with inversion of intervening material).

The Y chromosome also has medical importance in the field of transplantation. It has been known for almost 50 years that male tissue is occasionally rejected when grafted to a female recipient (Eichwald & Silmser, 1955). The male-specific factors responsible for this histoincompatibility has been termed H-Y antigen. There has been a focus on the Y chromosome in several species as a potential source of antigens that would not be present in females. To date, three Y genes have been immunologically identified as contributing to the H-Y antigen: SMCY (Wang *et al.*, 1995), USP9Y (Vogt *et al.*, 2000a), and UTY (Vogt *et al.*, 2000b).

Mapping on the Y Chromosome

The Y chromosome has been of interest to biologists for nearly a century, but it is only in the last few decades that the Y chromosome has been investigated on the genomic level. For many years the Y has been ignored or actively marginalized by the biological community as a chromosome with little impact on human biology, functioning only as a degenerate vessel to contain the testis determining factor (TDF). This activity, arguably the best known on the chromosome, is encoded by SRY and was only localized and cloned within the last 15 years (Page *et al.*, 1987; Sinclair

et al., 1990). There were, however, early indications that the Y played a more complex role in human biology. As the implications of X inactivation became known, the phenotypes in Turner Syndrome could then be ascribed to loss of factors present either on the X *or* Y chromosomes. Further, it was eventually determined that the TDF was not a component of the H-Y antigen, indicating that further translated factors must exist on the Y chromosome (Simpson *et al.*, 1987).

Further exploration of the chromosome would require genomic mapping tools, however. On other chromosomes, linkage mapping may be performed by correlating map distance to how often two markers are separated by recombination during meiosis (Ott, 1999). The majority of the Y chromosome is recombinationally isolated, however - only the very distal tips, termed the pseudoautosomal regions, pair and recombine with the X chromosome. The remainder of the chromosome is inherited from the father as a unit, changing only in the event of mutation. Occasionally, however, such mutations take the form of large-scale deletions, frequently resulting in the loss of a terminal portion of the chromosome, extending from within the chromosome to the very distal end of one of the arms. Given markers within the Y chromosome, these deletions can be used to order the markers with regard to one another. Such deletion mapping is accomplished by testing for the presence or absence of each marker on a panel of DNA from individuals with diverse deletions. Various marker orders are then considered, and the number of internal breaks (that is, regions with absent markers flanked by regions that contain present markers) are counted. The marker order with the fewest internal deletions is chosen as the "best" order, on the assumption that such deletions are rare compared to terminal deletions.

The first such map was constructed in 1986 from 23 markers in 27 deleted patients (Vergnaud *et al.*, 1986). The markers had been derived from randomly selected genomic fragments that had been identified as deriving from the Y chromosome. Testing was performed by Southern blotting digests of patient DNA, and probing with radiolabeled marker DNA. This procedure is tedious and not well suited to large-scale mapping. In 1992 a second-generation map employing 132 markers and 96 individuals was constructed (Vollrath *et al.*, 1992). This map was constructed employing the polymerase chain reaction (PCR), which greatly simplified scoring for the presence or absence of genomics markers.

Expansion of the map would still require the development of further markers, and the discovery of additional patients with unique breakpoints. The Page Lab has maintained a stock of cell lines and DNAs derived from patients with a variety of disorders, including sex-reversed females and infertile males. In many such cases Y chromosome deletions had been previously identified cytologically, and could then be more precisely characterized by marker content typing. As evidence grew that Y chromosome deletions might be responsible for some forms of male infertility, karyotypically normal infertile males began to be screened for smaller deletions by PCR, a process that identified new breakpoints useful for mapping (Kamischke *et al.*, 1999; Ma *et al.*, 2000; Reijo *et al.*, 1995).

The order of markers that had no breakpoint between them remained uncertain, however, and the discovery of new breakpoints in patients was slow. A second approach was then undertaken to generate additional breakpoints, this time in the form of genomic subclones. A yeast artificial chromosome (YAC) library was screened with existing Y markers to identify clones derived from the Y chromosome. These clones were then tested for

marker content by PCR. The results allowed both an ordering of the clones along the chromosome, as well as an increased resolution in the ability to order the markers relative to each other (Foote *et al.*, 1992). In this sense, each YAC terminus can be thought of as a new breakpoint that may be integrated with the patient deletion panel, once the YAC has been localized with respect to previously mapped markers.

Radiation Hybrid Mapping

Radiation hybrid (RH) mapping is an elegant technique pioneered in the last decade by David Cox and other researchers (Cox *et al.*, 1990). In order to begin mapping, a panel of hybrid cell lines must first be constructed. Figure 1 presents a simplified schematic for the steps involved. A cell line from the organism of interest is lethally irradiated with X-rays, causing the genome to fragment into many smaller pieces. This dying population of 'target' cells is then induced to fuse with a vast excess of healthy hamster cells which lack the thymadine kinase (TK) gene. If fusion occurs between a hamster cell and a target cell, the hamster genome will 'adopt' random fragments from the target genome into its own nucleus. The fraction of the target genome thus adopted, referred to as the retention frequency (RF), tends to vary around 10-30%, but is relatively consistent for hybrids within a particular panel. The retention frequency can then also be interpreted as the probability that any given fragment will be retained (adopted) by a particular hybrid. Target cells that fail to fuse will die as a result of the irradiation, and unfused hamster cells are eliminated by growth in HAT medium, which eliminates all but the hybrid cells which have incorporated the TK gene from the target cell line (this has the minor side effect of making map construction immediately around the TK locus impossible).

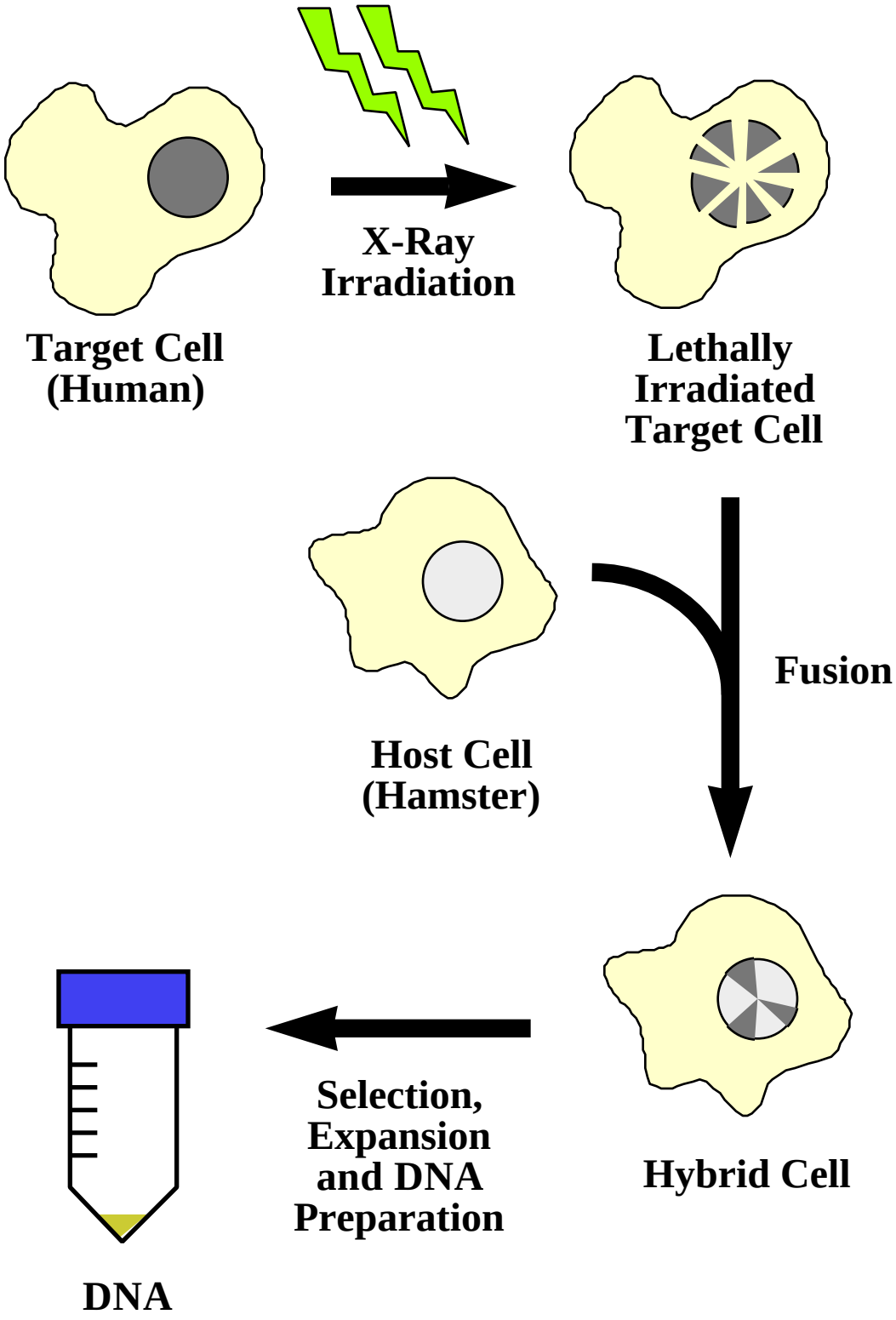


Figure 1. Schematic view of preparation of RH panel

Individual hybrid cells are then picked after selection and expanded in very large culture volumes. This culture is then used to generate a single, large preparation of DNA. The adopted target fragments are not irreversibly integrated into the hamster genome, and may be randomly lost during expansion of the culture. For this reason, while the DNA preparation will contain a uniform representation of a complete hamster genome, the adopted fragments will be present only at equal or lower stoichiometry than the hamster DNA - if an adopted fragment was lost early in the expansion, sequences contained within it will be present at much lower molarity than hamster sequences, or adopted sequences that were maintained during most of the expansion. This unpredictable non-clonal growth prevents the panel from being regenerated from frozen cell lines - each panel may be produced only once, hence the need to generate very large amounts of DNA in a single expansion. Typically on the order of 96 hybrids are expanded, as this quantity will conveniently fit on a single polymerase chain reaction (PCR) plate.

Once the panel DNAs have been generated, their application is very simple. A small quantity of each hybrid DNA is transferred to a 96-well plate (typically, some wells are reserved for positive and negative control samples as well). A PCR primer pair representing a sequence-tagged site (STS) marker is then added to each well, along with the remaining reagents necessary for PCR. After thermal cycling, each sample is then analyzed to determine if the base sequence for the STS was amplified in that particular hybrid. For RH projects on a heroic scale, such as performed at the Whitehead Genome Center, this detection step might be performed by a highly-automated fluorescent hybridization technique (Hudson *et al.*, 1995). More typically, however, the samples will be run out by agarose gel electrophoresis, and amplified DNA of the appropriate size detected by

ethidium bromide staining. It is worth emphasizing that RH mapping produces "plus / minus" data - each hybrid is assayed to determine simply if the STS base sequence was present (i.e. an amplified band was generated) or absent (no amplification). This contrasts to linkage mapping, where the markers must distinguish between two or more different alleles on recombining chromosomes, typically by slight differences in the size of the amplified product.

The data for each STS is compactly represented by a binary vector where each digit represents a particular hybrid. Digits are set to 1 if the corresponding hybrid was positive for the STS, or 0 if the hybrid was negative. Typically, the PCR for each STS will be repeated for the entire panel. If the second analysis of a hybrid generates a result that conflicts with the previous analysis (i.e. one test was positive, while the other was negative), then this ambiguous result is designated as 2 in the vector.

Distance between markers may now be determined by comparing any two vectors. In general, hybrids that differ between STSs (one being 0, the other 1) are interpreted as reflecting some physical separation between the two markers - the more such differences there are, the farther apart the markers must be. To understand the rationale behind this assumption, consider the situations that will lead to two markers being simultaneously present or absent in a hybrid cell. Let's consider an extreme case, where the two markers are absolutely linked (that is, with an effective distance of zero between them - perhaps they represent primer pairs that are just shifted by 20bp with regards to each other). In such a case, the hybrids between such markers will always agree, barring experimental or user error. If one marker has been retained on a fragment by a hybrid cell, then the other marker must also be present, and comparison between the vectors will both show '1' for

that hybrid. On the other hand, if one is lost because the fragment bearing it was not retained by a hybrid, then the other, being absolutely associated, must also have been lost on the same fragment, and both vectors will indicate '0' at the appropriate position.

As the distance between a pair of markers increases, however, the probability of an X-ray event inducing a break between them also increases (this assumes that the probability of an X-ray break occurring in a region is strictly related to the size of that region, an assumption that is generallyvalid). If such a break occurs, the two markers are now physically separated and on separate fragments. In this case, the retention or loss of each fragment will occur independently, and the markers will associate at random. The greater the distance, then, the greater the likelihood of break-induced separation, and the more likely that comparison of hybrids in a vector pair will observe a '1' for one STS and a '0' for the other.

For example, consider the following three vectors (where the '0' results have been faintly colored for increased legibility):

```
Marker A    000100000001000000001000000001000000001000000000...
Marker B    000100000001000000101000000001000000001000000000...
Marker C    000000010000000110000000100000100000010000000010...
```

Without resorting to mathematics, it is clear that Marker A is in close proximity to B, given that there are very few differences between the two vectors (only one hybrid differs in the part of the data shown). Conversely, Marker C is clearly at a much greater distance from A and B, as it has a very divergent vector (there is only one location in the visible data that shares a positive with A and B).

Of course, marker association is not typically determined by visual inspection. The vectors may be processed algorithmically to provide a numerical estimate of the distance between them. A theta score (θ) is an

estimation of the probability of an X-ray induced break occurring between a pair of markers during generation of the panel. Theta scores of 0 then indicate that the markers are never separated and are, to the resolution available from the panel, absolutely linked. A theta value of 0 would be calculated whenever a pair of vectors are identical. On the other extreme, theta values of 1 indicate that the markers are always separated, and therefore absolutely unlinked. This value would be returned when any correlations observed in a pair of vectors is at or below that expected from random chance. In practice, theta scores of 0.6 or less are assumed to represent reliable linkage between a pair of markers.

A theta score may be directly converted to a centiray (cR) measurement. Centirays are the moral equivalent of their counterpart in the linkage mapping world, the centimorgan. A centiray corresponds to the physical distance over which the probability of X-ray breakage is 1%. The actual distance spanned by a centiray will thus be dependent on the intensity of radiation used to generate a panel. When high radiation intensities are used, breaks are much more frequent than a panel exposed to low dosages, and the physical distance corresponding to a centiray is thus shorter. A centiray unit therefore is usually subscripted with the number of rads used during panel generation, for example cR_{6000}. Radiation dosage is carefully controlled during panel construction to determine the size of fragments generated. Low dosage panels, resulting in large fragments, are typically used for generating maps that can provide reliable linkage over long genomic distances with few markers. Higher dosage panels require more markers to produce reliably linked contigs, but allow the markers within the contigs to be ordered with respect to one-another at greater resolution.

The relationship between centirays and theta scores is quite straightforward:

$$cR = -100 \times \ln(1-\theta) \quad [1]$$

Centirays, in turn, may be converted directly to physical distance by multiplication with a simple conversion factor. To determine this factor it is necessary to know both the theta scores as well as the actual, physical separation between a set of markers. Like centimorgans, centirays and theta values provide reasonable estimates of whether markers are "near" or "far", but are unreliable for determining precise distances, particularly at the extreme limits of linkage. The errors in these calculations are most apparent if the distances between a series of markers are summed in an attempt to measure the size of a large region. Further, while gross deviations in the relationship between centirays and physical distance are uncommon, the conversion factor will vary depending on the region of the genome being assayed. The predominate cause of this uncertainty is likely due to a limit on the resolution of RH mapping due to "rounding errors" (given a finite set of hybrids, the precise value for a region can not be determined). However, it is likely that the predisposition towards X-ray breakage differs slightly in various regions of the genome, which will have the effect of skewing the conversion from one region to another.

The likelihood of linkage may also be computed from a vector pair. In such a case, the probability of the markers being linked is compared to the probability that the markers are unlinked, and a LOD score is returned. LOD scores are, by nature, on a logarithmic scale, so each increment indicates a 10-fold change in probability, with larger LOD scores indicating higher likelihoods of linkage. For both theta and LOD scores, the algorithms used will depend on the structure of the genome being analyzed - in particular,

standard algorithms have been developed for mapping either haploid or diploid genomes. The mathematics used in these calculations are presented by Jones (Jones, 1996).

Ultimately, one will wish to use the map data to generate a reliable order for the markers being tested. The simplest method for generating such an order is by distance minimization. The goal in such a scheme is to find a linear order such that the sum of the distance between all neighbors is the smallest possible total distance when all orders are considered. This task is identical to the traveling salesman problem - given a set of n cities, with known distances between each, what is the route the salesman should take that will touch on each city only once while traveling the shortest possible distance?

When n is very small, it is possible to exhaustively calculate all possibilities, and find the absolute best order (the one producing the absolute minimum distance). However, the number of possible orders increases as a function of n!, and the computational load increases at a slightly greater rate (since more elements need to be manipulated). For this reason, large data sets can not be exhaustively ordered. Fortunately, there are algorithms that allow rapid determination of orders with very low distances. It is impossible to prove that the distances reached by such methods are in fact the global minimum, but empirical evidence suggests that they are very close to it, if not in fact at the minimum.

An excellent algorithm termed "random cost" has been described by Wang and colleagues and is used within our lab for distance minimization (Wang *et al.*, 1994). The algorithm requires a table listing the objects to be ordered and the distances between all possible pairs. For RH mapping, this will be the theta matrix, an array containing the theta scores between all

marker pairs. To conceptualize the operation of the algorithm, imagine that we are exploring the "energy surface" of all possible orders. This is a multidimensional surface (visualize it in three dimensions) where the "height" of the curve represents the summed distance between neighbors at that point. We will start by placing a ball randomly on the curve, and will attempt to "shake" the ball to the lowest point of the surface, which corresponds to the minimum total distance and the marker order represented at that point. The algorithm proceeds through the following steps:

1. Calculate the distance for a hundred random orders to estimate the amount of variance (V) in distance possible for the data set. Pick another random order to start with.
2. Pick two random points in the order, invert all markers between those points, and recalculate the total distance. If the distance has decreased, or has increased but is less than V, allow the change. This step corresponds to shaking the surface - we send the ball flying, and if it lands lower, we let it stay there. We also let it stay in a higher position if it is within the variance. This is to allow the ball to escape local minima - it may have fallen into a "valley" of relatively good order, but not the global minimum.
3. Decrease V very slightly, and repeat step 2. Keep cycling between 2 and 3 until V reaches zero. This slowly decreases the magnitude of tolerated increases in the total distance- we are shaking the surface with decreasing vigor.
4. Reset V to maximum, and repeat the 2-3 cycle: return to a vigorous shaking of the surface, with gradual attenuation. This larger 2-4 cycle is repeated 100,000 times.

5. Repeat the cycle through steps 1-4 until the same minimum distance is calculated four times in a row - hopefully this implies that the order at the global minimum has been reached, or at least an order very near to it.

While in theory this analysis can be applied to an entire set of markers, it is generally useful to first cluster the markers into groups with high linkage probability. This reduces the chance of placing unlinked markers in close proximity, and greatly reduces the amount of computation needed to perform the overall analysis. A LOD value is chosen as a cutoff, and all possible pairwise comparisons are performed with the data set. Markers are added to a cluster only if they pass the LOD cutoff to at least one other marker within the group. These linkage groups are then internally ordered. The order of the groups with respect to each other is uncertain unless external data exists to indicate the larger organization of the linkage groups. Ideally, sufficient markers will be available such that the region studied will form a single linkage group.

Radiation hybrid maps have now been constructed for numerous organisms, at varying scopes and resolutions. The most dramatic endeavors have been genome-wide RH maps, where the marker density has been such that blanket coverage of all chromosomes is possible. Such maps have been produced for human (Gyapay *et al.*, 1996; Hudson *et al.*, 1995; Stewart *et al.*, 1997) (at roughly 250kb and 1000kb resolutions), zebrafish (Geisler *et al.*, 1999), pig (Hawken *et al.*, 1999), mouse (McCarthy *et al.*, 1997), rat (McCarthy *et al.*, 2000), dog (Mellersh *et al.*, 2000) and cat (Murphy *et al.*, 2000), while panels are available for other organisms as well (Kiguwa *et al.*, 2000; Womack *et al.*, 1997). These efforts have used large-fragment panels, which allow reliable linkage across large distances, and thereby reduce the demand on the number of STSs needed to provide the necessary coverage.

The need to procure sufficient markers is usually the limiting factor in genome-wide maps. At first glance, marker development for radiation hybrid maps is simplified compared to linkage maps, as it is not necessary to find allelic differences between strains. Radiation hybrid markers frequently suffer from homology with the host genome, however. Because each hybrid contains a complete host genome (which is present in equal or greater stoichiometry than the target fragments), markers selected from target sequence will occasionally amplify a homologous sequence in the host. In such a situation, every hybrid will be scored as positive (if the homologous product is the same size as the target), and the vector will be useless. Aside from this difficulty in marker generation, the simple increase in the cost and labor needed to construct a dense map have restricted most genome-wide analyses to large fragment panels.

High-resolution RH maps have been constructed over limited regions of a chromosome, however. Typically, these maps are generated to support efforts to localize or clone a gene associated with a disease or phenotype, and involve a handful of markers. Examples include late-onset hearing loss (Morell *et al.*, 2000) (5 markers), familial spastic paraplegia (Paternotte *et al.*, 1998) (90 markers), and prostate cancer (Arbieva *et al.*, 2000) (39 markers).

Recently, the high-resolution TNG panel has been used to type over 35,000 markers selected from across the human genome (Olivier *et al.*, 2001). This map incorporates 116 markers on the Y chromosome in 14 contigs of LOD 3 or better. At the time of this writing, only 42 of these markers reported in the literature were publicly available for analysis. Analysis of this subset indicates that the markers are of low retention frequency, a standard practice to help avoid multi-copy STSs. As a result, large portions of the Y chromosome are not represented in the marker coverage. A search of

known Y sequence for the primer sites represented by these markers verifies that very large regions of the Y are entirely lacking in marker coverage.

Marker Choice - the Diploid Tyranny

As mentioned above, map construction is generally marker-limited. A mapping project will likely start with STSs based on known sequence. Such sequences will typically be culled from public databases, or from within the mapping lab. Additional sequences can be generated by subcloning and sequencing random genomic fragments. However, a criteria for all standard mapping projects is that a sequence be single copy within a genome (which will be two copies in a diploid organism).

Single copy markers are selected for several reasons. First, it simplifies the mathematics for calculating theta scores or linkage between markers. At the heart of any calculation is the probability that a pair of vectors will both be positive for a hybrid, both negative, or one positive and the other negative. These probabilities are altered as the number of copies of a marker increases - in particular, it becomes more likely that a hybrid will be positive (since more copies are present to be retained by the hybrid), and thus more likely that a pair of markers will be jointly positive by chance alone.

Second, multicopy markers can serve as "bridges" that will cause two isolated groups to become mathematically linked. As mentioned previously, marker clustering based on linkage is typically a preliminary step to ordering. Imagine two marker clusters, one on Chr1 and the other on Chr2, with each cluster containing a copy of a duplicated marker. This marker will show strong linkage to the Chr1 group, as well as to the Chr2 group, such that during clustering the two groups will be combined into a single cluster. Ultimately this will serve to show Chr1 and Chr2 as a single linkage group - clearly an outcome to be avoided.

Finally, if the difference in copy number between two markers is too great, the pair will be indicated as being unlinked even if linkage exists between some members of the two markers. A difference in copy number will guarantee an increased number of hybrid pairs that disagree, since the higher copy number vector will contain more positives than the marker with lower ploidy.

These considerations result in the elimination from marker pools of any STS that might be multicopy. When large-fragment panels are being used, this restriction is usually only a minor nuisance, as sufficient sequence complexity exists over very long stretches of DNA to allow selection of novel markers. However, parts of the genome contain dramatically reduced sequence complexity - of particular interest to this discussion, the Y chromosome contains very little truly single copy sequence. In these cases, elimination of multicopy markers will result in regions with no marker coverage at all. For this reason, the map of the Y chromosome required the inclusion of markers suspected or known to be present in multiple locations.

The following chapters will describe the methods used to generate a high-resolution radiation hybrid map spanning the Y chromosome, which was then used as a framework for sequencing the chromosome. First additional, low copy markers were developed by a novel clone-based subtraction technique. Those markers were then typed on the TNG radiation hybrid panel and organized to generate a high-resolution map spanning most of the chromosome. Finally, this map was used to guide the selection of BAC clones which ultimately served as the basis for sequencing the chromosome. In the process of generating the map, software tools were developed to determine the likelihood of linkage between markers without making assumptions as to

their copy number. The methods used to perform this analysis are discussed in a separate chapter.

References

Adler H. T., Nallaseth F. S., Walter G., and Tkachuk D. C. (1997). HRX leukemic fusion proteins form a heterocomplex with the leukemia-associated protein SET and protein phosphatase 2A. *Journal of Biological Chemistry* **272:** 28407-28414.

Arbieva Z. H., Banerjee K., Kim S. Y., Edassery S. L., Maniatis V. S., Horrigan S. K., and Westbrook C. A. (2000). High-resolution physical map and transcript identification of a prostate cancer deletion interval on 8p22. *Genome Res* **10:** 244-57.

Charlesworth B. (1996). The evolution of chromosomal sex determination and dosage compensation. *Curr Biol* **6:** p149-62.

Charlesworth B. (1998). Sex chromosomes: evolving dosage compensation. *Curr Biol* **8:** pR931-3.

Cox D. R., Burmeister M., Price E. R., Kim S., and Myers R. M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250:** p245-50.

Eberhart C. G., Maines J. Z., and Wasserman S. A. (1996). Meiotic cell cycle requirement for a fly homologue of human Deleted in Azoospermia. *Nature* **381:** p783-5.

Eichwald E., and Silmser C. (1955). Untitled. *Transplant. Bull.* **2:** 148–149.

Foote S., Vollrath D., Hilton A., and Page D. C. (1992). The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science* **258:** p60-6.

Geisler R., Rauch G. J., Baier H., van Bebber F., Brobeta L., Dekens M. P., Finger K., Fricke C., Gates M. A., Geiger H., Geiger-Rudolph S., Gilmour D., Glaser S., Gnugge L., Habeck H., Hingst K., Holley S., Keenan J., Kirn A., Knaut H., Lashkari D., Maderspacher F., Martyn U., Neuhauss S., Haffter P., and et al. (1999). A radiation hybrid map of the zebrafish genome. *Nat Genet* **23:** p86-9.

Gromoll J., Weinbauer G. F., Skaletsky H., Schlatt S., Rocchietti-March M., Page D. C., and Nieschlag E. (1999). The Old World monkey DAZ (Deleted in AZoospermia) gene yields insights into the evolution of the DAZ gene cluster on the human Y chromosome. *Hum Mol Genet* **8:** p2017-24.

Gyapay G., Schmitt K., Fizames C., Jones H., Vega-Czarny N., Spillett D., Muselet D., Prud'Homme J. F., Dib C., Auffray C., Morissette J.,Weissenbach J., and Goodfellow P. N. (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* **5:** p339-46.

Hawken R. J., Murtaugh J., Flickinger G. H., Yerle M., Robic A., Milan D., Gellin J., Beattie C. W., Schook L. B., and Alexander L. J. (1999). A first-generation porcine whole-genome radiation hybrid map. *Mamm Genome* **10:** p824-30.

Hudson T. J., Stein L. D., Gerety S. S., Ma J., Castle A. B., Silva J., Slonim D. K., Baptista R., Kruglyak L., Xu S. H., and et al. (1995). An STS-based map of the human genome. *Science* **270:** p1945-54.

Jakubiczka S., Schnieders F., and Schmidtke J. (1993). A bovine homologue of the human TSPY gene [published erratum appears in Genomics 1994 Jan 1; 19(1):198]. *Genomics* **17:** p732-5.

Jegalian K., and Page D. C. (1998). A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394:** p776-80.

Jones H. B. (1996). Pairwise analysis of radiation hybrid mapping data. *Ann Hum Genet* **60:** p351-7.

Kalmykova A. I., Shevelyov Y. Y., Dobritsa A. A., and Gvozdev V. A. (1997). Acquisition and amplification of a testis-expressed autosomal gene, SSL, by the Drosophila Y chromosome. *Proc Natl Acad Sci U S A* **94:** p6297-302.

Kamischke A., Gromoll J., Simoni M., Behre H. M., and Nieschlag E. (1999). Transmission of a Y chromosomal deletion involving the deleted in azoospermia (DAZ) and chromodomain (CDY1) genes from father to son through intracytoplasmic sperm injection: case report. *Hum Reprod* **14:** p2320-2.

Karashima T., Sugimoto A., and Yamamoto M. (2000). Caenorhabditis elegans homologue of the human azoospermia factor DAZ is required for oogenesis but not for spermatogenesis. *Development* **127:** 1069-79.

Kiguwa S. L., Hextall P., Smith A. L., Critcher R., Swinburne J., Millon L., Binns M. M., Goodfellow P. N., McCarthy L. C., Farr C. J., and Oakenfull E. A. (2000). A horse whole-genome-radiation hybrid panel: chromosome 1 and 10 preliminary maps. *Mamm Genome* **11:** p803-5.

Lahn B. T., and Page D. C. (1997). Functional coherence of the human Y chromosome. *Science* **278:** p675-80.

Lahn B. T., and Page D. C. (1999a). Four evolutionary strata on the human X chromosome. *Science* **286:** p964-7.

Lahn B. T., and Page D. C. (1999b). Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome [published erratum appears in Nat Genet 1999 Jun; 22(2):209]. *Nat Genet* **21:** p429-33.

Lupski J. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics* **14:** 417-422.

Ma K., Mallidis C., and Bhasin S. (2000). The role of Y chromosome deletions in male infertility. *Eur J Endocrinol* **142:** p418-30.

Mazzarella R., and Schlessinger D. (1997). Duplication and distribution of repetitive elements and non-unique regions in the human genome. *Gene* **205:** p29-38.

McCarthy L. C., Bihoreau M. T., Kiguwa S. L., Browne J., Watanabe T. K., Hishigaki H., Tsuji A., Kiel S., Webber C., Davis M. E., Knights C., Smith A., Critcher R., Huxtall P., Hudson J. R., Jr., Ono T., Hayashi H., Takagi T., Nakamura Y., Tanigami A., Goodfellow P. N., Lathrop G. M., and James M. R. (2000). A whole-genome radiation hybrid panel and framework map of the rat genome. *Mamm Genome* **11:** p791-5.

McCarthy L. C., Terrett J., Davis M. E., Knights C. J., Smith A. L., Critcher R., Schmitt K., Hudson J., Spurr N. K., and Goodfellow P. N. (1997). A first-generation whole genome-radiation hybrid map spanning the mouse genome. *Genome Res* **7:** p1153-61.

Mellersh C. S., Hitte C., Richman M., Vignaux F., Priat C., Jouquand S., Werner P., Andre C., De Rose S., Patterson D. F., Ostrander E. A., and Galibert F. (2000). An integrated linkage-radiation hybrid map of the canine genome. *Mamm Genome* **11:** p120-30.

Mittwoch U. (1992). Sex determination and sex reversal: genotype, phenotype, dogma and semantics. *Hum Genet* **89:** p467-79.

Morell R. J., Friderici K. H., Wei S., Elfenbein J. L., Friedman T. B., and Fisher R. A. (2000). A new locus for late-onset, progressive, hereditary hearing loss DFNA20 maps to 17q25. *Genomics* **63:** 1-6.

Murphy W. J., Sun S., Chen Z., Yuhki N., Hirschmann D., Menotti-Raymond M., and O'Brien S. J. (2000). A radiation hybrid map of the cat genome: implications for comparative mapping. *Genome Res* **10:** p691-702.

Ogata T., and Matsuo N. (1995). Turner syndrome and female sex chromosome aberrations: deduction of the principal factors involved in the development of clinical features. *Hum Genet* **95:** p607-29.

Olivier M., Aggarwal A., Allen J., Almendras A. A., Bajorek E. S., Beasley E. M., Brady S. D., Bushard J. M., Bustos V., Chu A., Chung T. R., Witte A. D., Denys M. E., Dominguez R., Fang N. Y., Foster B. D., Freudenberg R. W., Hadley D., Hamilton L. R., Jeffrey T. J., Kelly L., Lazzeroni L., Levy M. R., Lewis S. C., Liu X., Lopez F. J., Louie B., Marquis J. P., Martinez R. A., Matsuura M. K., Misherghi N. S., Norton J. A., Olshen A., Perkins S. M., Perou A. J., Piercy C., Piercy M., Qin F., Reif T., Sheppard K., Shokoohi V., Smick G. A., Sun W. L., Stewart E. A., Fernando J., Tejeda, Tran N. M., Trejo T., Vo N. T., Yan S. C., Zierten D. L., Zhao S., Sachidanandam R., Trask B. J., Myers R. M., and Cox D. R. (2001). A High-Resolution Radiation Hybrid Map of the Human Genome Draft Sequence. *Science* **291:** 1298-1302.

Ott J. (1999). Methods of Analysis and Resources Available for Genetic Trait Mapping. *The Journal of Heredity* **90:** 68-70.

Page D. C., Mosher R., Simpson E. M., Fisher E. M., Mardon G., Pollack J., McGillivray B., de la Chapelle A., and Brown L. G. (1987). The sex-determining region of the human Y chromosome encodes a finger protein. *Cell* **51:** p1091-104.

Paternotte C., Rudnicki D., Fizames C., Davoine C. S., Mavel D., Durr A., Samson D., Marquette C., Muselet D., Vega-Czarny N., Drouot N., Voit T., Fontaine B., Gyapay G., Auburger G., Weissenbach J., and Hazan J. (1998). Quality assessment of whole genome mapping data in the refined familial spastic paraplegia interval on chromosome 14q. *Genome Res* **8:** 1216-27.

Reijo R., Lee T. Y., Salo P., Alagappan R., Brown L. G., Rosenberg M., Rozen S., Jaffe T., Straus D., Hovatta O., and et al. (1995). Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. *Nat Genet* **10:** p383-93.

Rice W. R. (1987). Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* **116:** p161-7.

Salo P., Kääriäinen H., Petrovic V., Peltomäki P., Page D. C., and Chapelle A. d. l. (1995). Molecular mapping of the putative gonadoblastoma locus on the Y chromosome. *Genes, Chromosomes & Cancer* **14:** 210-214.

Saxena R., Brown L. G., Hawkins T., Alagappan R. K., Skaletsky H., Reeve M. P., Reijo R., Rozen S., Dinulos M. B., Disteche C. M., and Page D. C. (1996). The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nat Genet* **14:** p292-9.

Schellhas H. F. (1974). Malignant potential of the dysgenetic gonad: Part II. *Obstetrics and Gynecology* **44:** 455-462.

Scully R. E. (1970). Gonadoblastoma. A review of 74 cases. *Cancer* **6:** 1340-1356.

Shanske A., Sachmechi I., Patel D. K., Bishnoi A., and Rosner F. (1998). An adult with 49,XYYYY karyotype: case report and endocrine studies. *Am J Med Genet* **80:** p103-6.

Sharpe R. M. (1994). Regulation of Spermatogenesis. *In* "The Physiology of Reproduction" (E. Knobil, and J. D. Neill, Eds.), pp. 1363-1434, Raven Press, Ltd., New York.

Silber S. (2000). Evaluation and treatment of male infertility. *Clinical Obstetrics and Gynecology* **43:** 854-888.

Simpson E., Chandler P., Goulmy E., Disteche C. M., Ferguson-Smith M. A., and Page D. C. (1987). Separation of the genetic loci for the H-Y antigen and for testis determination on human Y chromosome. *Nature* **326:** 876-8.

Sinclair A. H., Berta P., Palmer M. S., Hawkins J. R., Griffiths B. L., Smith M. J., Foster J. W., Frischauf A. M., Lovell-Badge R., and Goodfellow P. N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346:** 240-4.

Singh R. P., and Carr D. H. (1965). The anatomy and histology of XO human embryos and fetuses. *Anat. Rec.* **155:** 369-384.

Stewart E. A., McKusick K. B., Aggarwal A., Bajorek E., Brady S., Chu A., Fang N., Hadley D., Harris M., Hussain S., Lee R., Maratukulam A., O'Connor K., Perkins S., Piercy M., Qin F., Reif T., Sanders C., She X., Sun W. L., Tabar P., Voyticky S., Cowles S., Fan J. B., Cox D. R., and et al. (1997). An STS-based radiation hybrid map of the human genome. *Genome Res* **7:** p422-33.

Tsuchiya K., Reijo R., Page D. C., and Disteche C. M. (1995). Gonadoblastoma: molecular definition of the susceptibility region on the Y chromosome. *American Journal of Human Genetics* **57:** 1400-1407.

Vergnaud G., Page D. C., Simmler M. C., Brown L., Rouyer F., Noel B., Botstein D., de la Chapelle A., and Weissenbach J. (1986). A deletion map of the human Y chromosome based on DNA hybridization. *Am J Hum Genet* **38:** p109-24.

Vogt M. H., de Paus R. A., Voogt P. J., Willemze R., and Falkenburg J. H. (2000a). DFFRY codes for a new human male-specific minor transplantation antigen involved in bone marrow graft rejection. *Blood* **95:** 1100-5.

Vogt M. H., Goulmy E., Kloosterboer F. M., Blokland E., de Paus R. A., WillemzeR., and Falkenburg J. H. (2000b). UTY gene codes for an HLA-B60-restricted human male-specific minor histocompatibility antigen involved in stem cell graft rejection: characterization of the critical polymorphic amino acid residues for T-cell recognition. *Blood* **96:** 3126-32.

Vollrath D., Foote S., Hilton A., Brown L. G., Beer-Romero P., Bogan J. S., and Page D. C. (1992). The human Y chromosome: a 43-interval map based on naturally occurring deletions. *Science* **258:** p52-9.

Wang W., Meadows L. R., den Haan J. M., Sherman N. E., Chen Y., Blokland E., Shabanowitz J., Agulnik A. I., Hendrickson R. C., Bishop C. E., and et al. (1995). Human H-Y: a male-specific histocompatibility antigen derived from the SMCY protein. *Science* **269:** 1588-90.

Wang Y., Prade R., Griffith J., Timberlake W. E., and Arnold J. (1994). A fast random cost algorithm for physical mapping. *Proc. Natl. Acad. Sci. USA* **91:** 11094-11098.

Western P. S., Harry J. L., Graves J. A. M., and Sinclair A. H. (2000). Temperature-dependent sex determination in the American alligator: expression of SF1, WT1 and DAX1 during gonadogenesis. *Gene* **241:** 223–232.

Womack J. E., Johnson J. S., Owens E. K., Rexroad C. E., Schlapfer J., and Yang Y. P. (1997). A whole-genome radiation hybrid panel for bovine gene mapping. *Mamm Genome* **8:** 854-856.

Chapter 2

Generation of low copy-number Y chromosome markers through YAC subtraction

Charles Tilford, Steve Rozen, Helen Skaletsky, Tomoko Kawaguchi, Michael Rosenberg, David Page

Introduction

A problem frequently encountered when initiating a mapping project is insufficient markers to provide reasonable coverage of the region of interest. New markers can be generated by sequencing random genomic fragments, but this method will produce more markers from outside the area than within. If large clones are available from a region, these may be subcloned from shotgun libraries, providing sequence known to arise from the region of interest. Even this technique is inefficient, however, as large tracts of the genome contain genome-typical repeats, such as SINEs and LINEs, and are inappropriate for mapping purposes. On the human Y chromosome, this problem is exacerbated by the presence of Y-specific repeated sequences, leading to most randomly chosen Y markers being present at multiple dispersed locations.

While the Y chromosome does contain genome-typical repeats, it also contains multiple repeated sequences of a much more complex structure. These Y-specific amplicons are quite large, with repeat units ranging from tens of kilobases to megabases in size. The repeats often contain intact and transcribed genes, which in some cases have been demonstrated to function in spermatogenesis (Reijo *et al.*, 1995). Conservation between repeat units typically exceeds 99%, and in extreme cases can be as high as 99.99%. Some repeats, such as the TTY family of genes, are dispersed across the

chromosome. Others, including TSPY, are present as large tandem arrays. The most dramatic structures, however, are inverted repeats organized as huge palindromes, exemplified by the DAZ cluster on the long arm of the chromosome (Kuroda-Kawaguchi *et al.*, 2001). Taken together, these repeated sequences occupy a great majority of Y sequence.

To address the problem of generating new, low-copy markers in such a landscape, a novel subtraction approach was undertaken. A chromosome-spanning contig of yeast artificial chromosomes (YACs) had been constructed using previously developed Y markers. A subset of 57 of these clones was chosen that formed a minimum tiling path across the chromosome. These YACs were then fragmented and ligated to adapters, such that the fragmented pools could be maintained and manipulated by PCR. A subtraction protocol was then applied individually to each pool, using that pool as tracer, and all other pools from non-overlapping YACs as drivers. In this manner, not only genome-typical repeats were subtracted, but Y-specific repetitive sequences as well.

These subtracted pools were then subcloned and sequenced, with the intent of using the sequences to develop sequence tagged sites (STSs). These markers were then used to construct the radiation hybrid map described later (Tilford *et al.*, 2001). The process of mapping the markers served as the most informative analysis of the success of the subtraction, and has validated the technique as a robust method for generating single or low-copy markers.

Materials and Methods

Yeast artificial chromosomes and subtraction pools. YACs were chosen from the 1992 Y map (Foote *et al.*, 1992) such that complete coverage of the chromosome was accomplished with a minimum number (57) of clones (Figure 1). DNA was purified from each YAC, and digested

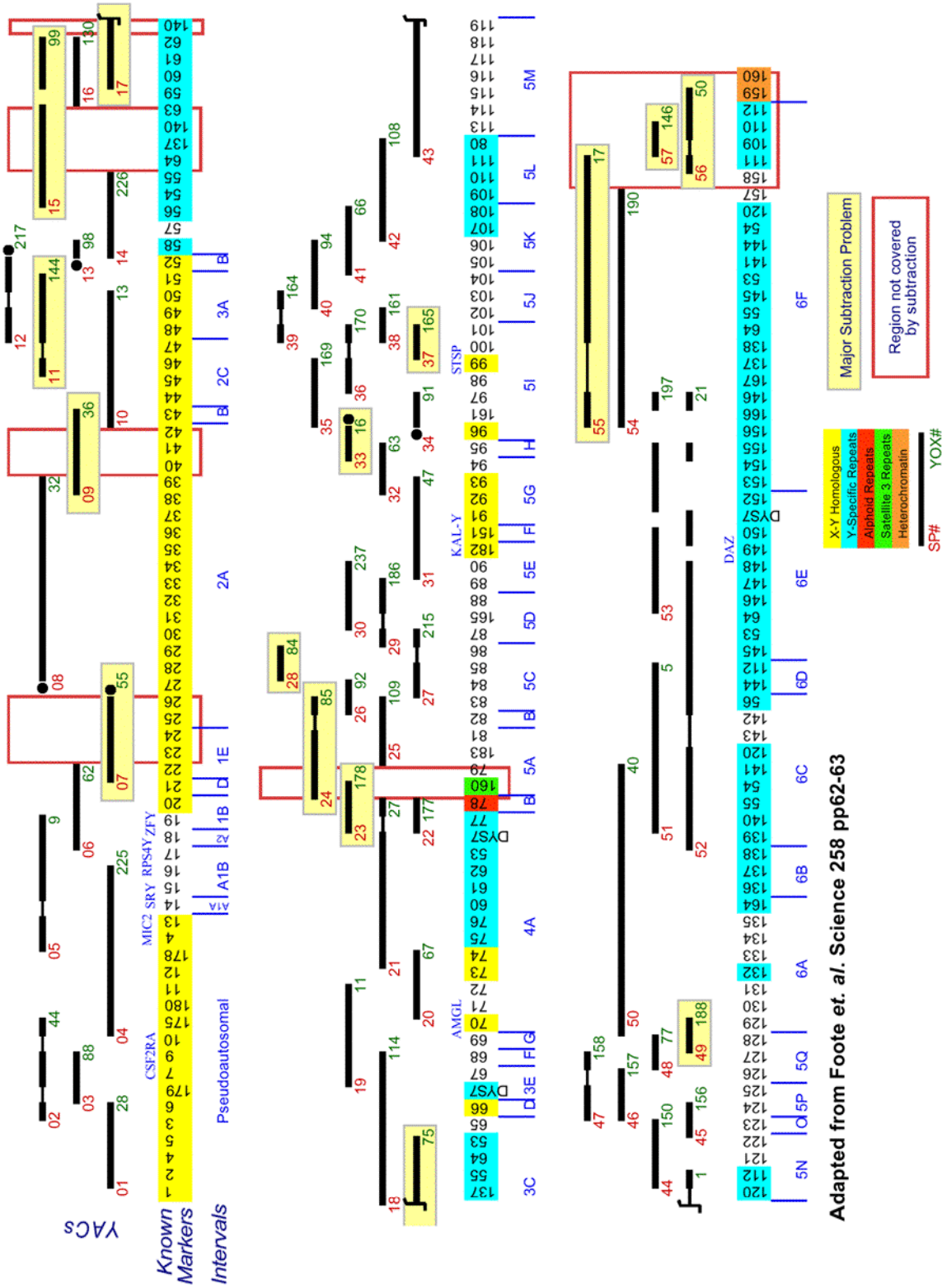


Figure 1. YACs used in subtraction protocol

with Sau3A (recognition site GATC). This produced a pool of fragments with predicted average size of 256bp. Adapters were ligated to these fragments to allow uniform amplification of the pool by PCR. Several adapter sequences were used, such that driver pools would use a different adapter than tracer pools. This would allow amplification of subtracted tracer without amplifying any trace contaminants from the driver population.

Subtraction. Driver material was created by separately amplifying each YAC with biotinylated adapters. For each tracer YAC, a driver pool was created by combining driver material from all YACs that did not overlap with the tracer. Tracer and driver were then denatured for 3 min at 100°C and then combined at a 1:100 ratio and allowed to hybridize overnight at 65°C. The reaction was then bound to avadin-coated beads (Dynal) and incubated for 10 minutes at room temperature. The reaction was then spun through a filter, allowing separation of the biotinylated driver and associated sequences from the subtracted tracer, which remains in solution. The isolated tracer was then reamplified and subjected to three more rounds of subtraction with fresh driver.

Subcloning. Subtracted tracer pools were digested with Sau3A and spun through Sephacryl S-300 (Amersham Pharmacia Biotech) to remove adapters. These fragments were ligated into pBluescriptII KS+ (Stratagene) and introduced into DH5-α cells by electroporation. Transformations were selected on ampicillin plates, and 48 colonies were picked from each of the 57 subtractions. Colonies were grown in 130µl LB/ampicillin in 96-well microtiter plates overnight at 37°C. Plates were incubated in an airtight container containing standing water to provide a water-saturated atmosphere and thus prevent evaporation. After growth, 130µl of 80% glycerol was added to each well, and the plates were stored at -80°C.

Sequencing. A master mix containing 1x Promega PCR buffer (50mM KCl, 10mM Tris-HCl pH 9.0, 0.1% Triton X-100, 1.5 mM $MgCl_2$), 0.1mM each dNTP, 1µM each M13F (GTAAAACGAACGGCCAGT) and M13R (AACAGCTATGACCATG) primers, and 0.05U/µl Taq polymerase was prepared, with 20µl being added to each well of a 96-well plate. A "hedgehog" with 96 thin pins (designed for accessing 384 well plates) was then inserted into a partially thawed bacterial subclone stock plate, and used to "inoculate" the plate containing PCR master mix. A small quantity of bacteria, bearing the plasmid of interest, was thus transferred to each well. Thermal cycling (94°C for 5min, 20 cycles of [94°C for 30sec, 55°C for 30sec, 72°C for 3min], 72°C for 1 hour) then lysed the bacteria and amplified the insert carried by the plasmid. The samples were purified by spinning through Sephacryl 300. Cycle sequencing reactions contained 4µl of this purified PCR product, 0.1µl nested vector primer (CGAATTCCTGCAGCCCGGGGGA) and 3.325µl dye terminator mix (Perkin Elmer). The reactions were cycled (35 cycles of 96°C for 30sec, 50°C for 15sec, 60°C for 4min), spun through Sephadex G50-50 (Amersham Pharmacia Biotech), dried for an hour at 60°C, resuspended in 7µl formamide and denatured at 95°C for 3 minutes. The samples were loaded onto acrylamide gels and sequenced on an ABI370A (Applied Biosystems).

Sequence analysis and primer choice. Sequences were self-compared using BLAST to identify similar sequences. Sequences containing an internal Sau3A site were judged to be chimeric and were split into two separate sequences. The best quality representative from each sequence grouping was then submitted to Primer3 to select primers for forming an STS (Rozen & Skaletsky, 1997). Parameters were set to select primers of about 20bp and 50% GC content, annealing temperature near 58° but within 1°C of the other

primer in the pair, and product sizes of 100-200bp (smaller product sizes were preferred in order to minimize the chance of choosing primers on either side of an unrecognized chimera junction). When optimal conditions could not be found, the program would find pairs with the closest match to those specified. Primers were ordered at 20µM (Research Genetics). Each STS was tested on male and female DNA to determine the quality of the assay and identify male specific markers. Markers producing visible bands of the appropriate size were then tested on hamster DNA and a mixture of human and hamster DNA. Those assays which produced an identifiable human band were then selected for typing on radiation hybrid panels. Markers were generally confirmed as Y-derived by demonstrating linkage to known Y markers; in several cases, however, typing on a monochromosomal panel (Coriell - NIGMS Human/Rodent Somatic Cell Hybrid Mapping Panel #2) was used to determine the chromosomal source of the marker.

Results & Discussion

Initial analysis of subtraction subclones, based on the size of the inserts, indicated that the subtraction had produced a significant diversity of cloned sequences. This observation was confirmed upon sequencing; of 2634 samples sequenced, a total of 921 sequence groups were assembled. Of these, only 16 represented known repeats, suggesting that the subtraction had been highly successful in removing genome-typical repetitive sequences. Roughly two-thirds of the markers were found to produce male-specific bands; Figure 2 shows a representative gel of the assay used to determine male specificity. The remainder were still considered for radiation hybrid testing, however, as significant portions of the Y chromosome are known to share homology with the X chromosome, and in some cases autosomes. A total of 227 markers

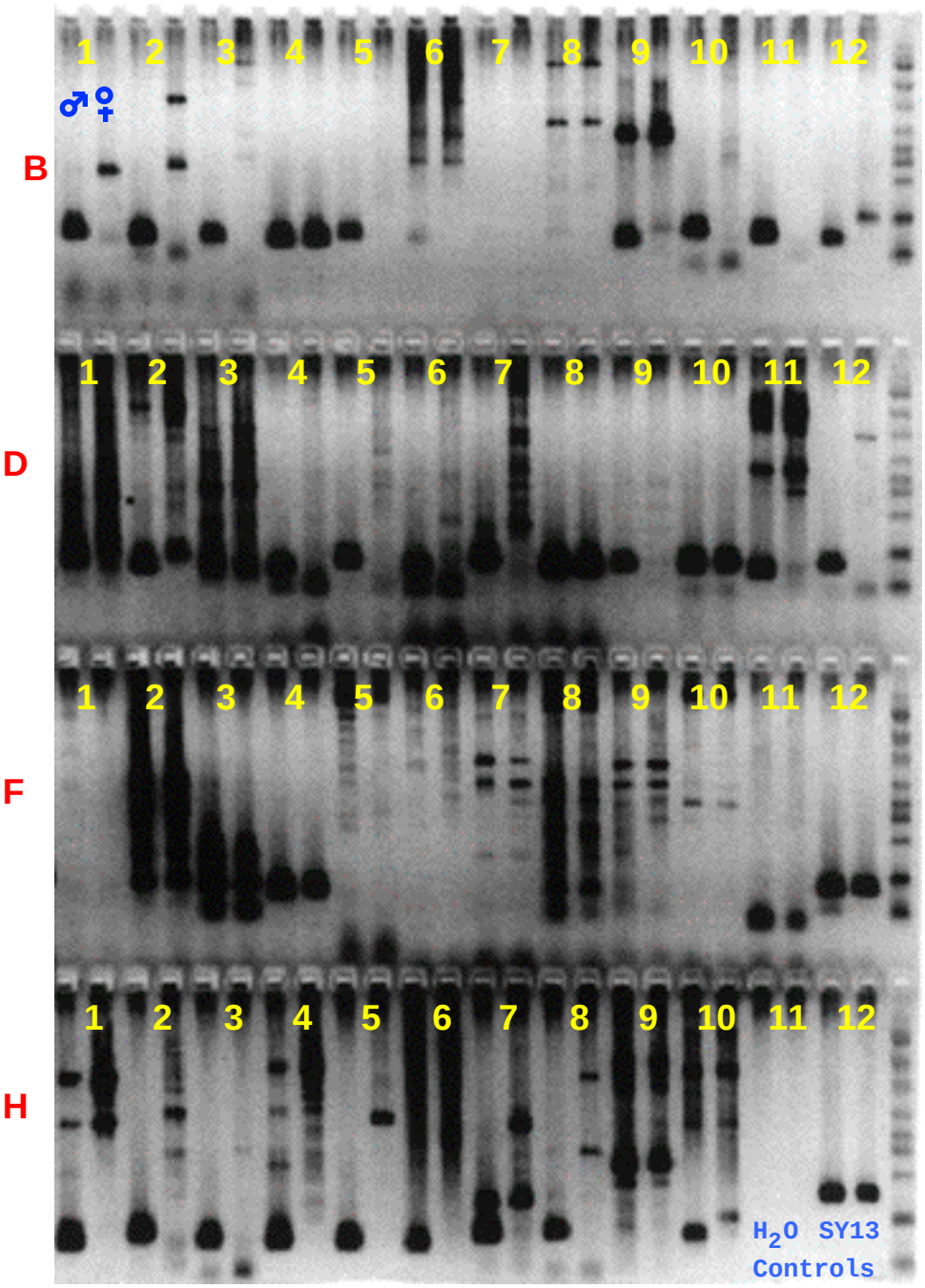


Figure 2. PCR trial of 46 subtraction markers on male and female DNA. Each marker is tested in a pair of assays, with 46,XY normal male DNA as template for the left assay, and 46,XX normal female DNA for the right. Markers recognizing only Y sequences amplify only in the male (e.g. B5), while some recognize other sequences in the genome and amplify in the female as well. Some of these markers are still identifiable as containing Y-specific sequences as the male band is distinct from the female (e.g. B1), while others produce no Y-specific bands (e.g. B4). A few assays fail to amplify altogether (e.g. B7) or amplify too many sequences to be useful (e.g. D3). Gel image is inverted for legibility.

were later found to derive from the X or from autosomes, but not have any homology to the Y.

These non-Y markers are almost certainly derived from chimeric portions of the YACs used in the subtraction. A known limitation of YAC clones is their tendency to incorporate multiple, unlinked fragments of DNA (Haldi *et al.*, 1994). Detection of such material is difficult at best, as YACs are generally characterized only by the presence or absence of markers suspected of being near the presumed location of the YAC. If a YAC was chimeric for non-Y chromosomal sequences, these would go undetected during the initial screens of the library, as markers recognizing these arbitrary sequences were not tested. During subtraction, however, these sequences would be greatly selected for, as they would be composed of unique sequences not present in the Y chromosome segments represented in the driver pool. A total of 14 subtraction pools derived from YACs known to contain Y sequence produced only autosomal markers, and thus did not contribute to the final map. Six additional pools generated significant numbers of autosomal markers, while also providing some useful Y chromosome STSs. Taken together, these results would indicate that at least 35% of the 57 YACs used in this experiment where chimeric.

Markers were also characterized on human and hamster DNA to verify that the primer pair would not amplify hamster sequences. If the underlying sequence for the STS was significantly homologous to sequences present in the hamster, then amplification of the ubiquitous hamster material could result in the inability to observe amplification (or more appropriately, lack of amplification) from the human target sequence. A representative gel from this screen is shown in Figure 3. While it was not uncommon to observe ubiquitous bands in all hamster-containing hybrids, they typically were at a

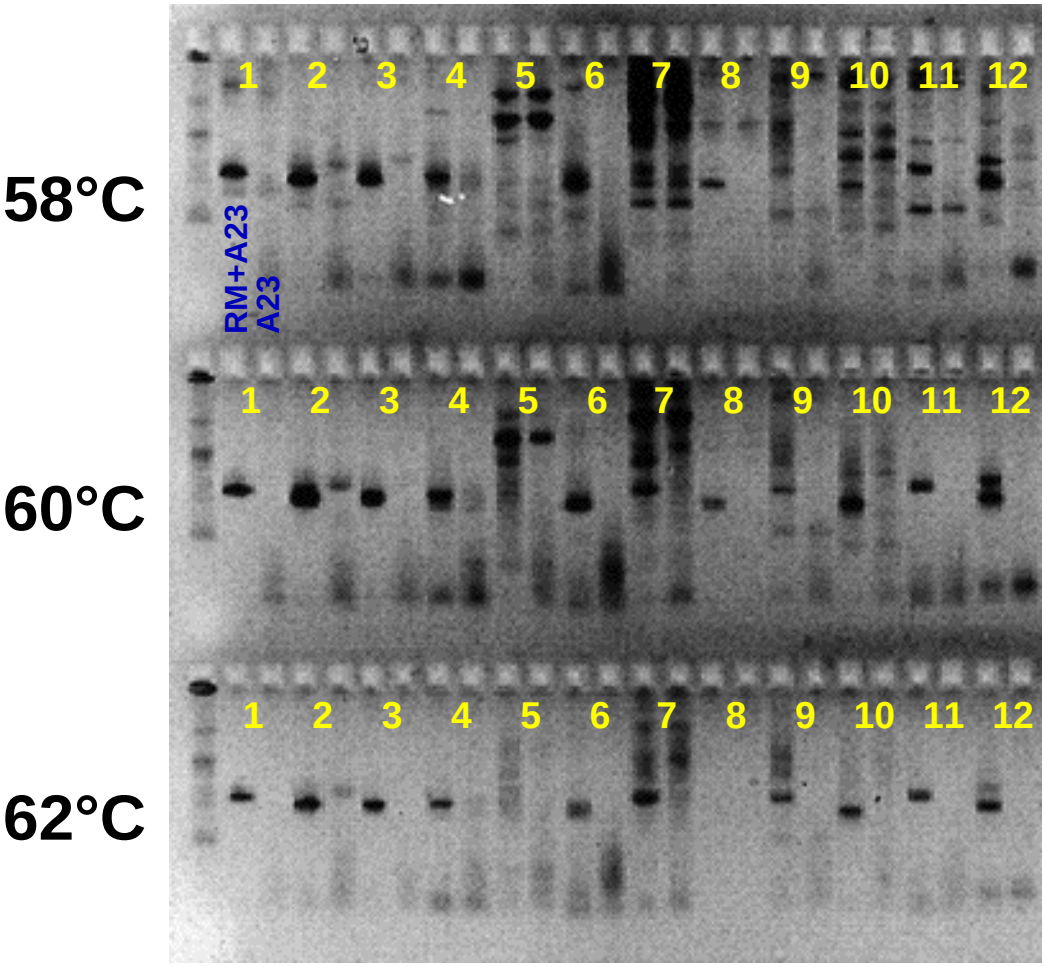


Figure 3. PCR trial of 12 subtraction markers at different temperatures on human and hamster DNA. Each marker is tested in a pair of assays, with a mixture of human and hamster DNA (RM+A23) used as template in the left assay, and hamster DNA alone (A23) as the right. These two samples would represent the possible template conditions encountered in a radiation hybrid panel. Each assay is also tested at three annealing temperatures, with the hope that at least one temperature will be high enough to provide discrimination against potential hamster targets, but low enough to still bind to the human target. In most cases the desired human sequence is amplified at all temperatures, while the hamster only sample remains unamplified. Some markers require the lower annealing temperature to amplify (e.g. 8), whereas others require a higher temperature to eliminate unwated hamster products (e.g. 7). Yet others fail at all temperatures (e.g. 5). Gel image is inverted for legibility.

greater, or occasionally lower, size than that expected, and thus did not present a problem in marker typing.

There are then two major limitations to this procedure for generating novel markers. The first is the identification and organization of sufficient numbers of large genomic clones to use as material for tracer and driver. In this case, the YACs had been previously identified in a separate mapping project. For a new organism, construction and screening of a YAC library, followed by mapping of identified YACs, would represent a significant investment of resources. Second, chimerism significantly limits the number of YACs that provide useful markers. Given that a minimum tiling path is used, elimination of a whole YAC results in a large hole in marker coverage.

Even given these challenges, in this situation the technique proved exceptionally useful in providing adequate coverage of the Y chromosome to begin sequencing. A total of 340 subtraction STSs ultimately proved informative on the radiation hybrid map. Combined with 350 previously identified markers, this proved sufficient to blanket most of the chromosome at a resolution of roughly one marker per 50kb, sufficient to identify a near-complete BAC contig for sequencing.

Conclusions

It is very unlikely that this technique will be applied again for the purpose of developing markers for genomic mapping. Given advances in sequencing technology and changes in the methodologies being used, it is probable that higher organisms will be sequenced starting directly from BAC libraries (for high quality whole-genome sequence) or by whole-genome shotgun approaches (for rapid assessment of commercially valuable regions). Smaller genomes will be exclusively shotgun sequenced. In none of these cases is a map essential.

However, there may be situations where the methodology described here could be useful. Unless a radical breakthrough in sequencing speed and cost is achieved, it will be a long time before every microorganism is sequenced. In particular, for many single-celled organisms a representative strain from each species will be sequenced, with the remaining strains being skipped in favor of more divergent genomes. The YAC subtraction technique could prove useful in identifying sequence differences between strains of the same species. Many bacterial genomes are comparable in size to a single YAC (although a bacterial genome is generally significantly more complex than a stretch of mammalian DNA of the same length). In this case, one strain would serve as driver, the other as tracer. This subtraction technique could then represent a relatively rapid, low-cost method for identifying sequence differences between bacterial strains. This would have particular utility in determining virulence factors differing between harmless and pathogenic strains.

References

Foote S., Vollrath D., Hilton A., and Page D. C. (1992). The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science* **258:** p60-6.

Haldi M., Perrot V., Saumier M., Desai T., Cohen D., Cherif D., Ward D., and Lander E. S. (1994). Large human YACs constructed in a rad52 strain show a reduced rate of chimerism. *Genomics* **24:** 478-84.

Kuroda-Kawaguchi T., Skaletsky H., Minx P., Brown L., Rozen S., Wilson R., Waterston R., and Page D. (2001). The AZFc region of the human Y chromosome: a complex of massive amplicons, testis-specific gene families, and uniform recurrent deletions. *Unpublished.*

Reijo R., Lee T. Y., Salo P., Alagappan R., Brown L. G., Rosenberg M., Rozen S., Jaffe T., Straus D., Hovatta O., and et al. (1995). Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. *Nat Genet* **10:** p383-93.

Rozen S., and Skaletsky H. J. (1997). Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Tilford C. A., Pooler L., Skaletsky H., Rozen S., Kawaguchi T., and Page D. C. (2001). Radiation hybrid mapping of multi-copy markers on the human Y chromosome. *pre-submission - Chapter 3 in this document*.

Chapter 3

Radiation hybrid mapping of multi-copy markers on the human Y chromosome

Charles A. Tilford, Loreall Pooler, Helen Skaletsky, Steve Rozen, Tomoko Kawaguchi, David C. Page

Abstract

Many genome mapping techniques consistently utilized on autosomes or the X chromosome fail when applied to the Y chromosome. Genetic mapping may only be applied near the telomeres, as there is a lack of recombination across the bulk of the chromosome. Radiation hybrid mapping traditionally requires uniform coverage of a region with single-copy markers, but such sequences represent the minority on the Y. We report here the generation of a high-resolution radiation hybrid map spanning the chromosome, and including markers of varying copy number. The incorporation of multi-copy markers required the map to be constructed in segments, with linkage and ordering of markers determined locally, and the resulting contigs then assembled into the larger whole. This map provided the framework for the sequence-ready BAC map previously reported (Tilford *et al.*, 2001). The map encompasses 617 markers in 28 contigs, which were successfully ordered in all but the most repetitive of regions.

Introduction

In the past decade, whole genome radiation hybrid mapping has evolved into an indispensable tool for genomic mappers (Cox *et al.*, 1990). By utilizing a collection of hybrid cell line DNAs, each containing a random fraction of the genome of interest, a researcher may very quickly determine

linkage within a set of sequence tagged sites (STSs), and ultimately order these markers with respect to one another. Provided with one of many whole genome panels, the researcher must only select or generate sufficient STSs to cover the region of interest, then test each against the panel's DNAs by PCR to determine if the STS target sequence is present or absent in each hybrid. The resulting data may then be analyzed with multiple algorithms to determine the extent of linkage between the STSs, and ultimately their relative order within the genome.

There are some constraints applied to the selection of the STS markers to be used. In particular, the analysis of the data is greatly simplified if each STS is present only once within the genome. Repetitive sequences are present throughout the genomes of higher organisms, and pose two difficulties in map construction. First, markers derived from these regions tend to result in inappropriate linkage between distant portions of the genome - if a marker is present in two otherwise unlinked regions of the genome, it will serve as a "bridge" between those areas, which will appear linked through the repeated marker. Second, information about specific copies is lost. Typing of results indicates only that *at least* one copy is present in the hybrid; data on precisely how many copies are present in a positive hybrid, or which of the specific copies they might be, are lost.

Fortunately these problems are relatively easy to avoid in most mapping projects. The majority of the genome contains sufficient single-copy sequences to allow suspected multicopy markers (either because they produced an abnormally high fraction of PCR-positive hybrids, or show sequence homology to known repeats) to be discarded, and the region to be represented instead by a nearby single-copy sequence. In this fashion several

radiation hybrid maps of the human genome have been successfully constructed (Gyapay *et al.*, 1996; Hudson *et al.*, 1995; Stewart *et al.*, 1997).

Conditions on the human Y chromosome are not conducive to standard mapping techniques, however. Radiation hybrids represent the most attractive high-resolution mapping option available, given that most of the chromosome lacks recombination and therefore can not be linkage mapped. This non-recombining region (NRY) is a very poor substrate for developing single-copy STSs, however. Like the remainder of the genome, the NRY contains genome-typical repeats, in particular LINEs. In addition, however, the chromosome also contains Y-specific amplicons. Unlike the "junk" DNA associated with genome-typical repeats, these amplified regions occupy large, continuous portions of the NRY (often in excess of 100kb) and contain actively transcribed genes. Homology between repeats generally exceeds 99%. Some of these repeats are organized in simple tandem arrays, such as the TSPY cluster. Others are structured as huge inverted repeats, forming palindromes that can span megabases (Kuroda-Kawaguchi *et al.*, 2001a). In addition, some of these regions share close homology to other chromosomes, particularly the X.

In such regions it is impossible to provide uniform coverage of the chromosome with single-copy markers. In order to develop a high-resolution comprehensive map of the Y chromosome, it would then be necessary to include multicopy markers. The most extensive radiation hybrid map of the human genome contains over 36,000 markers, including 116 on the Y chromosome (Olivier *et al.*, 2001). This map utilizes the TNG panel, which provides very high resolution of marker ordering, at the expense of requiring a large number of markers to assure linkage across a chromosome. Analysis of the publicly available markers indicates that they are universally of low

retention frequency, however. It was suspected that these markers would then fail to cover high-copy regions of the Y chromosome. Electronicidentification of STS binding sites within Y chromosome sequence confirms that large portions of the Y are not represented by these markers. Such avoidance of high retention frequency markers is advisable elsewhere in the genome, but on the Y chromosome it greatly restricts the scope of mapping.

The markers chosen for the map presented here were selected from a range of sources, and were rejected only if the assay was extraordinarily repetitive, or of very poor PCR quality. By allowing consideration of high-copy markers, it was hoped that greater coverage of the chromosome could be obtained. Inclusion of these markers required a more careful analysis of the data, however. Most radiation hybrid mapping projects begin by first performing an analysis of the data to determine clusters of markers that are linked at high likelihood. Applying such a global analysis to data containing multicopy markers will result in inappropriate linkage of distant groups. In would then be necessary to perform linkage analysis only on subsets of markers known to be in general proximity to each other. Clearly, this would be very difficult in the absence of any additional information; fortunately, low-resolution maps developed prior to this project provided coarse orientation for most of the markers utilized in the study. Use of these maps, which have an average combined resolution of about 250kb, can then allow the assembly of the higher resolution (estimated at 50kb) TNG map, even in regions of high repeat content. We discovered in the course of the project that the TNG panel was not only ideal for its high resolution, but because its relatively low retention frequency allows useful mapping data to be collected for repetitive markers as well.

Materials and Methods

Marker characterization. YAC subtraction was used to generate 921 new sequence tagged sites (STSs) for use in Y chromosome mapping (Tilford *et al.*, 2001). In addition, about 200 previously developed Y markers were available, and over 200 BAC end markers were developed during the construction of the BAC map. The markers were then tested against both 25 ng A23 hamster DNA, and 25 ng A23 hamster DNA mixed with 25 ng RM human DNA (Research Genetics). Reactions were in Promega buffer (500 mM KCl, 100 mM Tris-HCl pH9, 1% Triton X-100, 15mM $MgCl_2$) with 1U Taq polymerase in a total volume of 20µl. PCR conditions were 4 min 94°C, 35 cycles of (94°C for 30 sec, annealing temperature for 2 min, 72°C for 2 min) and 72°C for 7 min. The annealing temperature was set to 2°C higher than the temperature predicted by Primer 3 (Rozen & Skaletsky, 1997), usually corresponding to 58-64°C. Markers with distinct human products were then advanced to radiation hybrid typing.

Radiation hybrid typing. A Packard MultiProbe 204DT liquidhandling robot was used to dispense each of the 90 TNG hybrids (Research Genetics), plus 6 control samples (water, A23 hamster DNA, two samples of RM human DNA, one RM human sample at 1:5 dilution and one RM human sample at 1:20 dilution) to a 96 well plate. All samples were delivered in 7µl volume, comprising 25ng DNA in the undiluted samples. After dispensing, the plates were allowed to dry overnight, and then stored at room temperature for up to 30 days. The robot was then used to add 10µl diluted primer pair (corresponding to 15 nmol for each primer) to each well, and the plate again dried overnight. Once dried, 15µl PCR mix (1x Promega buffer, 100 nM each dNTP, 1U Taq polymerase) were added to each well, the wells were covered with 15µl mineral oil (Sigma #M3516), and placed on thermal

cyclers. Cycle conditions were as above. After cycling, 7µl loading buffer (75% glycerol, 3mM EDTA) were added to each sample and the reactions were run on 2% SeaKem ME (BioWhittaker) agarose gels.

Gels were visualized by ethidium bromide staining and recorded with a CCD image capture system (Alpha Inotech). Gel images were converted to GIF format and stored on a UNIX file server. Images were scored using the JavaGel applet (all software generated for use in map construction is available at `http://www.mit.edu/people/ctilford/CAT.html`) to generate vectors (a numerical representation of the PCR state for each hybrid) for each STS.

Results

Over 1200 markers were ultimately tested on the TNG panel,including previously characterized Y markers (Foote *et al.*, 1992; Vollrath *et al.*, 1992), markers derived from the YAC subtraction, and markers developed from BAC end sequence during construction of the BAC map (Tilford *et al.*, 2001). Linkage to chromosomes other than the Y were determined for 227 of the subtraction markers, which were then excluded from further analysis These markers likely arose from chimerism in the tracer YAC during subtraction, as such sequences would be unique when compared to driver.

The distribution of retention frequencies for the markers that were ultimately placed on the map is shown in Figure 1. Markers derived from both the subtraction and BAC ends tend to have lower retention frequencies than previously developed Y markers ("Other Y STSs"). This likely reflects the success of the subtraction in removing repetitive sequences. BAC end markers, on the other hand, were developed late in the sequencing project with the goal of identifying new BACs to close gaps in the physical map. As such, they were electronically screened against known human sequence, including the Y, and have been biased towards unique sequences

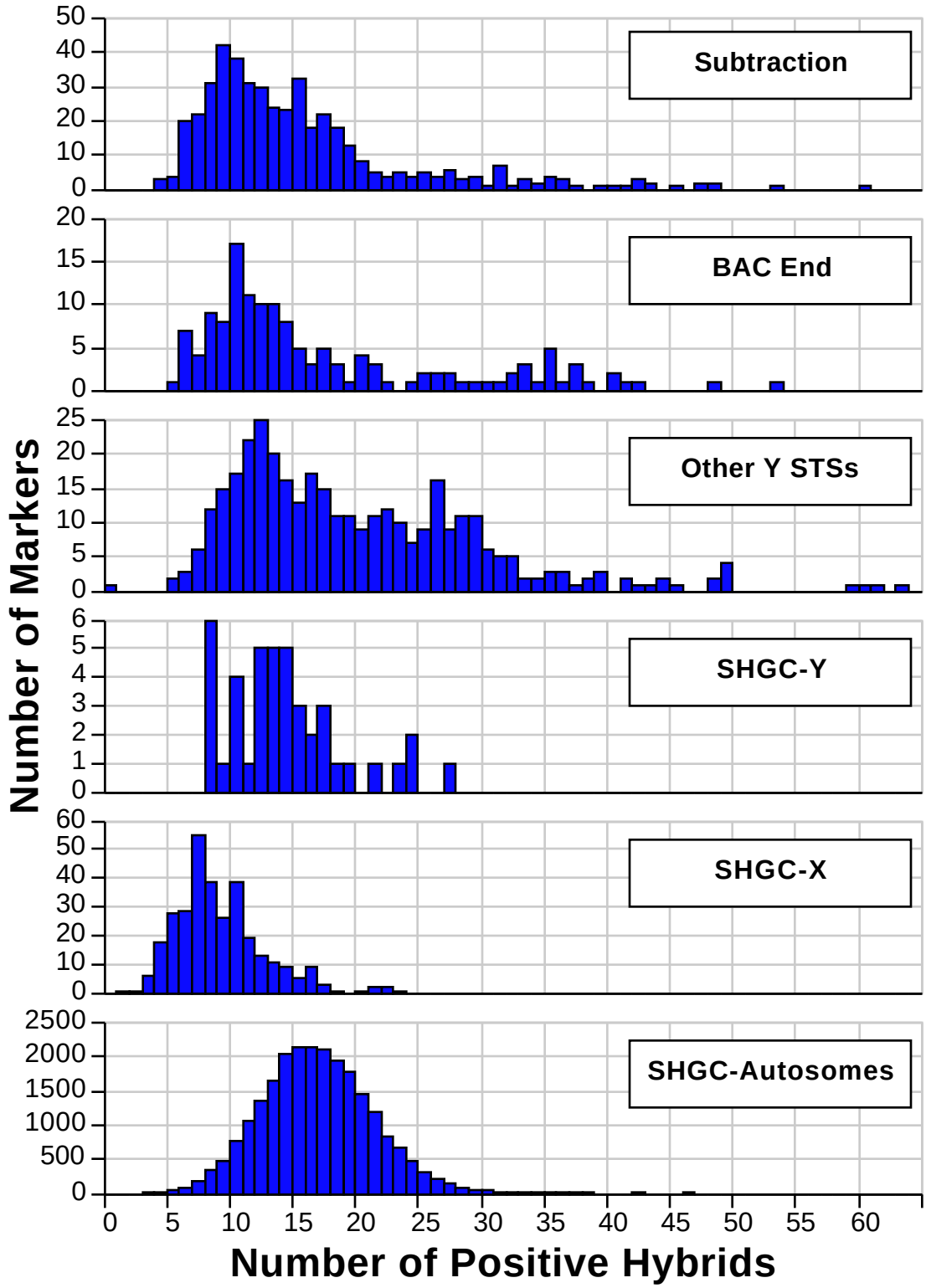


Figure 1. Retention frequencies of markers mapped on the TNG RH panel. The distribution of number of positive hybrids for several marker sets are shown. Top three panels this study, bottom three Stanford Human Genome Center.

By comparison, the markers used by the Stanford Human Genome Center have much lower average retention frequencies. Markers from the two haploid chromosomes (X and Y) rarely exceed 25 positives, and even the vast majority of STSs positioned on autosomes score less than 30 positives.

Initial gross assignment of marker position was then determined by the source of the marker. A majority of previously developed markers had been mapped either on a panel of patient DNAs with partially deleted Y chromosomes (Vollrath *et al.*, 1992), or on a physical map of Y chromosome YACs (Foote *et al.*, 1992). These markers had preliminary positions assigned consistent with these two maps. The subtraction protocol generated markers derived from individual clones in the YAC map, such that these markers could in a similar fashion be assigned to the preliminary positions based on the location previously determined for each tracer YAC. Finally, BAC end markers were developed, typically from BACs that had already been oriented on the growing BAC map. The preliminary positioning of these markers was then made such that they would be consistent with the position determined for the source BAC. Marker placement was also aided by the calculation of theta scores, the probability of a break occurring between a pair of markers (Jones, 1996). A visual representation of the theta matrix, the two-dimensional array representing the theta scores for all possible pairwise comparisons, was consulted to aid in initial placement of newly typed markers.

Once markers had been roughly placed, they were then analyzed to find groups of markers linked at high likelihood. This analysis was performed for isolated clusters to prevent inappropriate linkage of distant, repeated markers. The scope of each cluster was expanded until questionable linkage was observed between regions known to be distant. For the majority of the chromosome, analysis was performed with LOD 5 stringency to produce 42

groups. Each marker in such a group would then be linked to at least one other marker with a likelihood of at least 10^5. Within the DAZ region on the distal long arm, however, this likelihood proved too low to produce reliable groups, as the entire region was grouped in a single cluster. A higher likelihood of 10^{10} was used in this region to produce 7 linkage groups. Each linkage group was then ordered by random cost analysis, which attempts to find the minimum sum of theta distances between neighbors (Wang *et al.*, 1994). Orientation of a linkage group to its neighbors could frequently be determined either by previously known map order, or by maximizing the likelihood of linkage between the terminal markers of the two linkage groups.

Figure 2 presents the final ordering of the markers. The theta values for all pairwise comparisons within 30 positions of each marker are indicated above the plot as a color gradient (see key). In general, theta values represented in color (below 0.6) represent likely linkage. Visual inspection of the theta matrix can then quickly identify regions of closely placed markers, as these form blocks of color. Unlinked markers will show gray values between them, with a gap being indicated by extension of a gray region down to the diagonal. The linkage likelihood between neighbors is indicated below each marker pair. Even when the linkage likelihood is below LOD 3, linkage between neighboring groups is often suggested by inspection of the theta matrix surrounding the region. Below the linkage likelihoods is a plot indicating the number of positive hybrids for each marker (indicated as black rectangles), on a linear scale ranging from 0 to 75 positive hybrids. The data is plotted on a series of yellow and white bands indicating the number of copies of the marker that would be most likely to produce the observed the number of positive hybrids. The lowest yellow band represents values most likely to derive from a single copy, with each yellow band above it

representing an additional two copies. This is a rough estimate only; actual copy number may vary significantly from this value.

Many of the markers shown in Figure 2 were also used to screen high-density BAC filters. The BACs identified were then assembled into a physical map for use as a substrate in sequencing the chromosome (Tilford *et al.*, 2001). The number of BACs that were positive for each of these markers is also plotted on the same scale (blue circles). A red triangle at the top of the scale indicates that more than 75 BACs were positive for that marker. Markers that were not used in the BAC screen will have only a black rectangle indicating number of positive TNG hybrids.

Inspection of the theta matrix in Figure 2 indicates that in nearly all cases random cost ordering of the markers has performed very well. Ideally, starting at the diagonal and visually following the theta matrix away from a particular marker at a 45° angle (that is, comparing the marker to progressively distant neighbors) should take the viewer smoothly down the theta scale. Initially values might start in the red - yellow part of the scale, indicating low theta values for nearby neighbors, and then proceed into the more distant green and blue theta scores, before finally entering gray regions of markers too distant to reliably link. This pattern is observed in almost all parts of the chromosome.

A notable exception is in the AZFc region, at the very distal end of the map. In this region, just above the area marked "DAZ" there is a cluster of marker pairs that show very low theta values (they are indicated in red and orange) but are interrupted by a block of tightly linked markers (the largest red cluster just above "DAZ"). The order presented here is the best produced through distance minimization, but has separated two groups of markers that should be adjacent to one-another.

This map wraps around over three pages. A small region of the map (15 markers, indicated by shaded regions) is duplicated when it crosses a page break. To find the theta value for a pair of markers, trace upwards at a 45° angle from each marker to find the diamond above and between the two markers. Linkage likelihoods are listed between each neighbor, and are listed in red when below LOD 3. Plot at bottom indicates number of positive hybrids (black rectangles) and positive BACs (blue circles) for each marker.

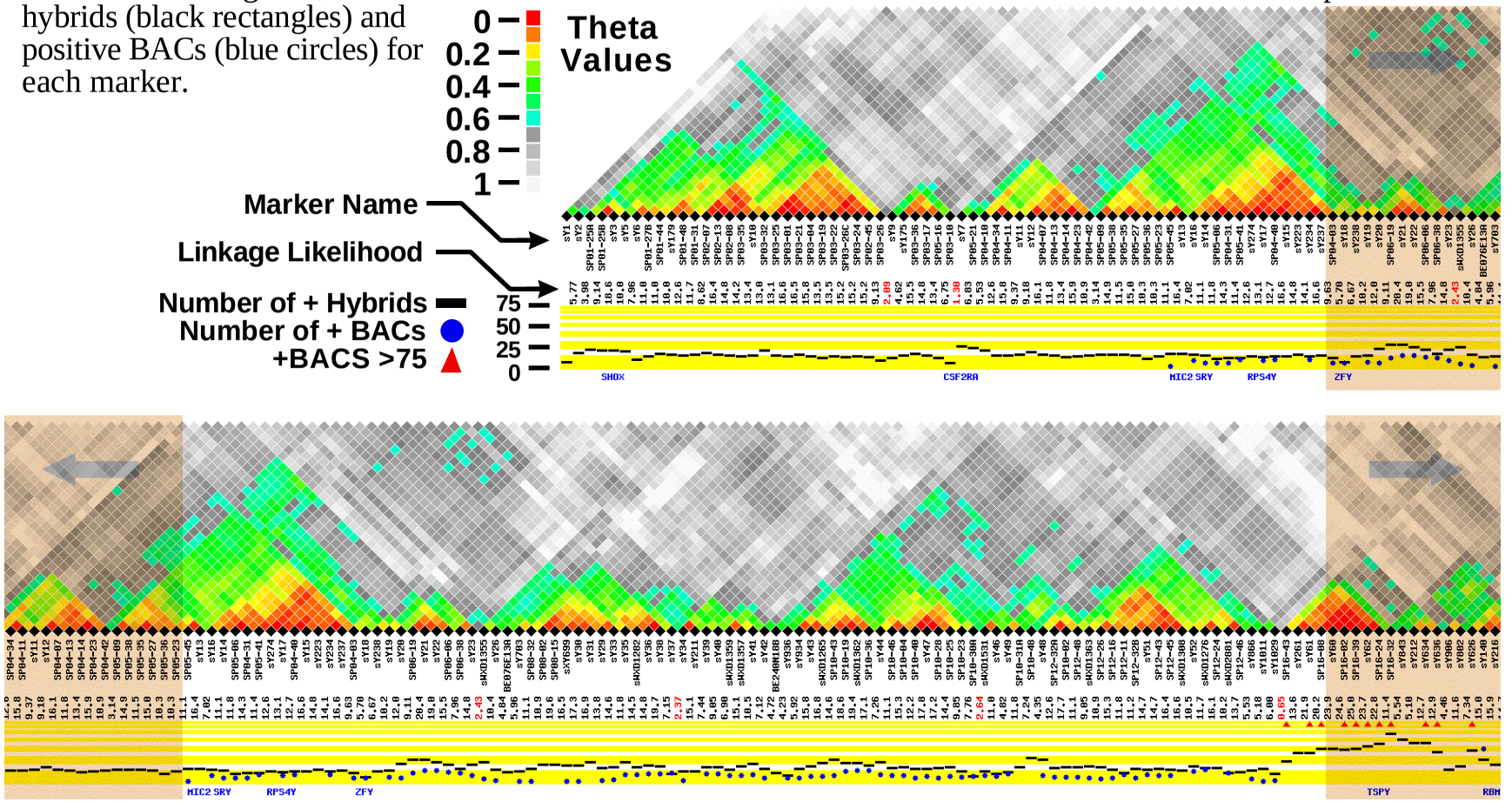


Figure 2. Radiation hybrid map of the Y chromosome **1 of 3**

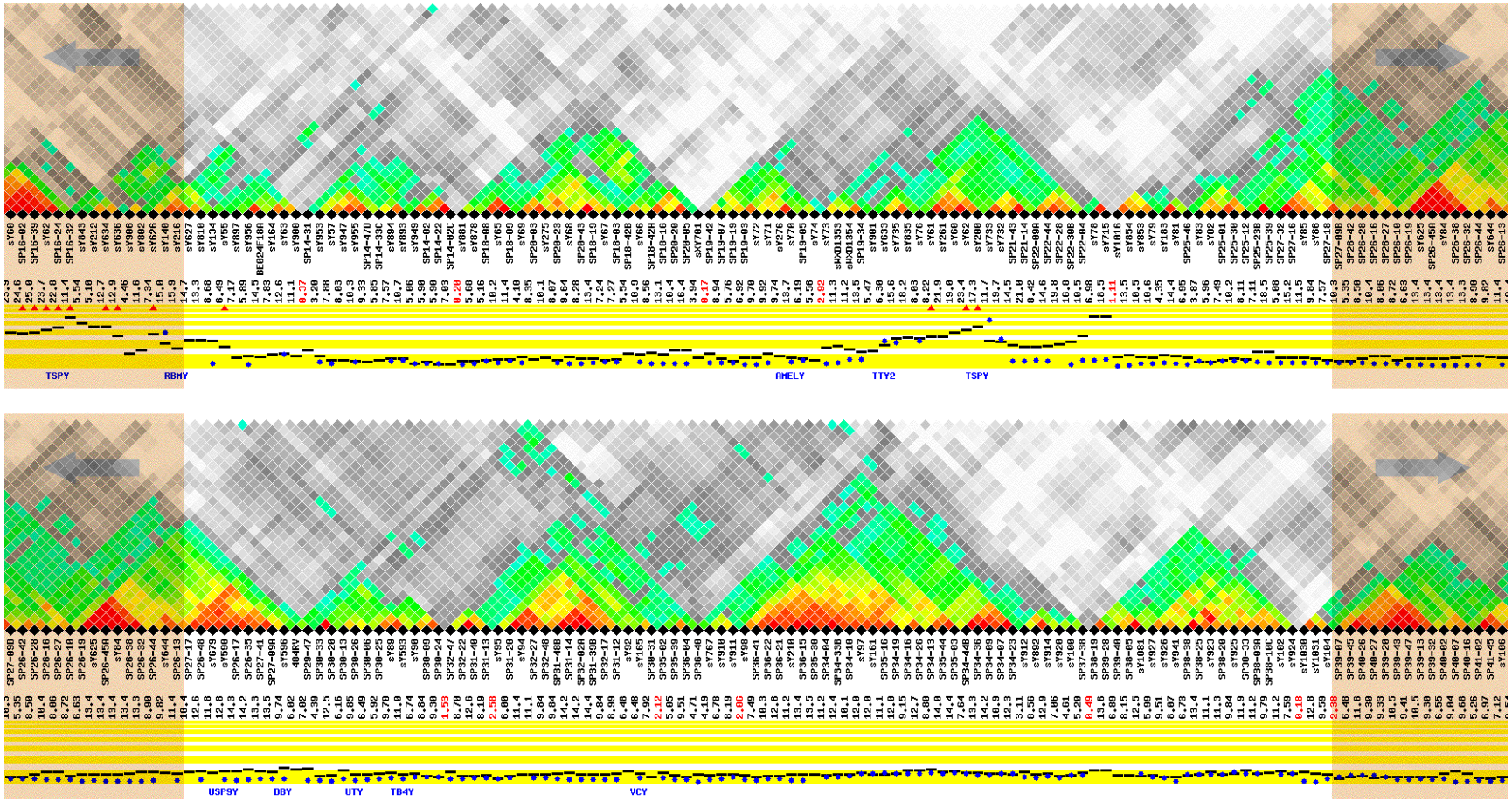


Figure 2. ***continued***

2 of 3

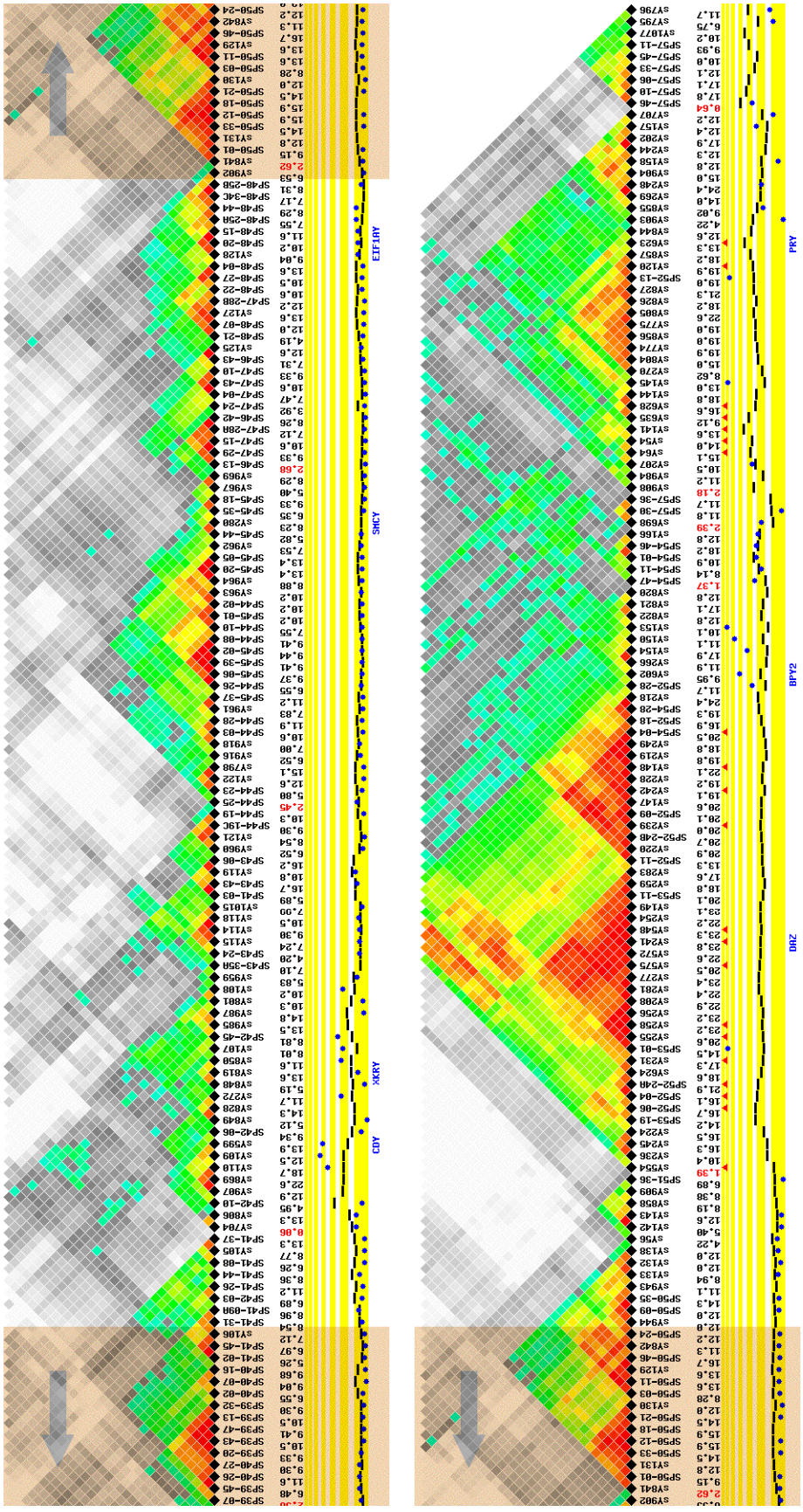


Figure 2. ***continued***

3 of 3

This map was produced to support the BAC map underlying the Y chromosome sequencing project. Once Y chromosome sequence was available(Kuroda-Kawaguchi *et al.*, 2001b), the radiation hybrid map was compared to it to measure the reliability of the map. Electronic PCR (ePCR) identified the location of likely STS sites within the sequence based on the proper orientation and spacing of primer sites. The order of these electronically identified markers was then compared to that determined by radiation hybrid mapping. Table 1 summarizes the results of this comparison. A total of 585 linked, neighboring markers in the RH map were also found in the sequence by ePCR. The number of intervening markers observed in the sequence was then determined, as well as the actual distance between the markers, as calculated from the sequence.

Discussion

The markers employed in the map span an extreme range of copy numbers, with single copy STSs at the lower end, and centromeric sequences at the upper. To provide complete coverage of the chromosome it was necessary to include the higher copy number markers in the map. Figure 1 provides a comparison between the markers used in this project with those chosen for the Stanford map. The Stanford mapping effort has used STSs clustered around the retention frequencies expected for a single copy marker. We have attempted to use similar low-copy markers wherever possible, but found it necessary to incorporate markers with retention frequencies over twice those utilized by Stanford in some circumstances.

A radiation hybrid panel provides the highest information content when exactly half of the hybrids are positive, as this maximizes the number of possible positive/negative patterns that could be generated for each marker. It would then seem that an ideal panel would have a retention frequency of 0.5,

# of intervening markers	Number of STS pairs	Average distance (kb)	Average Theta Score
0	182	19.3 ± 36.2	0.137 ± 0.118
1	96	29.4 ± 29.1	0.161 ± 0.138
2	45	43.5 ± 68.1	0.140 ± 0.124
3	35	31.6 ± 36.1	0.120 ± 0.124
4	34	50.0 ± 42.5	0.176 ± 0.127
5	17	73.2 ± 67.2	0.270 ± 0.238
>5	56	144 ± 164	0.324 ± 0.239

Table 1: Summary of comparison between RH map and sequence. Pairwise comparisons between 585 linked neighbors in the RH map. The number of markers found to intervene between the pair are listed in the first column. The second column indicates the number of pairs observed for that category of interventions. Fewer interventions indicates a more accurate ordering of the data. For each category, the average distance separating the pairs of markers in the sequence is indicated, as well as the average theta score calculated from the RH vectors.

and the observed value of 0.12 for TNG would appear low. For this project, however, it was very fortunate that the panel was constructed with a low retention frequency. Single copy markers in most instances proved to have an adequate number of positive hybrids for accurate map assignment. Markers with a large number of dispersed copies still produced intelligible vectors, however. An STS recognizing five unlinked sequences would be positive in roughly half the hybrids, while one recognizing 15 unlinked copies in the genome would be identified in about 85% of the hybrids (an information content comparable to that produced by a single copy marker). A panel with a retention frequency of 0.3 would provide more information for ordering single copy markers, but a 5 copy STS would be estimated to produce positive results in 83% of the hybrids. If an "ideal" panel with a single copy retention frequency of 0.5 was used, the same 5 copy STS would be found positive in 97% of the hybrids. Such a panel would provide very little useful information for positioning multi-copy markers.

By using previously existing, low-resolution maps, it was possible to coarsely organize most of the markers. Theta scores could then be used to refine the positions of neighboring markers. However, searching for linkage groups in the entire data set simultaneously consistently failed, as repetitive markers would link dispersed regions of the chromosome. As an example, markers derived from a YAC containing the minor TSPY repeat cluster would strongly link that region to the major TSPY cluster, roughly 6 Mb away. The scope of clustering was then limited to allow only local markers to be compared - if inappropriate linkage (as judged by previous map data) was observed, the scope of the region was further limited. In some cases, known repetitive markers could be removed, but in some portions of the

chromosome it was essential to retain repeated STSs in the analysis in order to obtain coverage in this high-resolution panel.

Random cost distance minimization of the theta scores proved to be a rapid and effective means of marker ordering. Comparison to sequence reveals that, within linkage groups, most markers have been ordered properly. Many of the mis-ordered marker pairs are within 50kb of one another, and thus near the predicted resolution of the panel. The order of the groups themselves is also in good concordance with the sequence, although six groups were oriented in the opposite direction relative to the sequence. It was discovered that the Y chromosome of the sequenced individual contains a short arm inversion relative to the Y present in the RH panel (and to most males typed within this lab) (Tilford *et al.*, 2001). This region of the reference sequence was thus electronically inverted in order to compare the sequence to the map.

One region proved exceptionally difficult to analyze, however. The AZFc region, located distal on the long arm just proximal of the heterochromatin, contains a highly repetitive region with striking organization. Spanning 4.5Mb, this part of the chromosome contains four very large repeat units, the largest being 1.5Mb, with homology between repeats generally exceeding 99.9%. The repeats are organized into two palindromic pairs, and contain several known genes (Saxena *et al.*, 2000). Markers within this area showed very strong association with one another, to the extent that more stringent linkage conditions were imposed to prevent association of all markers into a single group. Once assigned to isolated groups of LOD 10 linkage, the markers were reasonably ordered by distance minimization. However, it was impossible to determine the larger, palindromic structure of the region from the radiation hybrid data. That task would

ultimately require careful analysis of sequenced BACs to fully dissect the organization of the region.

Even within the DAZ region, the map served its primary function, to support construction of a sequence-ready BAC contig. Although the higher organization of the markers within the palindromes could not be determined, it was still possible to reliably discern clusters of nearby markers. These were then chosen as pools of geographically associated markers for screening of the BAC filters, greatly facilitating the process of identifying regionally-related BACs for the sequencing effort.

Completion of the map presented here indicates that radiation hybrid mapping within highly repetitive regions is feasible, but requires external information to accurately order repeat-containing regions. The coincident development of both the RH map and the physical BAC map was beneficial to both projects, as each served as a confirmation of the other, as well as providing marker resources useful to both. This strategy of simultaneous high-resolution radiation hybrid typing and BAC screening would be useful not only in multi-copy environments, but in any situation where a clone-based sequencing project was being undertaken.

References

Cox D. R., Burmeister M., Price E. R., Kim S., and Myers R. M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250:** p245-50.

Foote S., Vollrath D., Hilton A., and Page D. C. (1992). The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science* **258:** p60-6.

Gyapay G., Schmitt K., Fizames C., Jones H., Vega-Czarny N., Spillett D., Muselet D., Prud'Homme J. F., Dib C., Auffray C., Morissette J.,Weissenbach J., and Goodfellow P. N. (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* **5:** p339-46.

Hudson T. J., Stein L. D., Gerety S. S., Ma J., Castle A. B., Silva J., Slonim D. K., Baptista R., Kruglyak L., Xu S. H., and et al. (1995). An STS-based map of the human genome. *Science* **270:** p1945-54.

Jones H. B. (1996). Pairwise analysis of radiation hybrid mapping data. *Ann Hum Genet* **60:** p351-7.

Kuroda-Kawaguchi T., Skaletsky H., Minx P., Brown L., Rozen S., Wilson R., Waterston R., and Page D. (2001a). The AZFc region of the human Y chromosome: a complex of massive amplicons, testis-specific gene families, and uniform recurrent deletions. *Unpublished.*

Kuroda-Kawaguchi T., Skaletsky H., Minx P., Brown L., Rozen S., Wilson R., Waterston R., and Page D. (2001b). The DNA sequence of the human Y chromosome. *Unpublished.*

Olivier M., Aggarwal A., Allen J., Almendras A. A., Bajorek E. S., Beasley E. M., Brady S. D., Bushard J. M., Bustos V., Chu A., Chung T. R., Witte A. D., Denys M. E., Dominguez R., Fang N. Y., Foster B. D., Freudenberg R. W., Hadley D., Hamilton L. R., Jeffrey T. J., Kelly L., Lazzeroni L., Levy M. R., Lewis S. C., Liu X., Lopez F. J., Louie B., Marquis J. P., Martinez R. A., Matsuura M. K., Misherghi N. S., Norton J. A., Olshen A., Perkins S. M., Perou A. J., Piercy C., Piercy M., Qin F., Reif T., Sheppard K., Shokoohi V., Smick G. A., Sun W. L., Stewart E. A., Fernando J., Tejeda, Tran N. M., Trejo T., Vo N. T., Yan S. C., Zierten D. L., Zhao S., Sachidanandam R., Trask B. J., Myers R. M., and Cox D. R. (2001). A High-Resolution Radiation Hybrid Map of the Human Genome Draft Sequence. *Science* **291:** 1298-1302.

Rozen S., and Skaletsky H. J. (1997). Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Saxena R., de Vries J. W., Repping S., Alagappan R. K., Skaletsky H., Brown L. G., Ma P., Chen E., Hoovers J. M., and Page D. C. (2000). Four DAZ genes in two clusters found in the AZFc region of the human Y chromosome. *Genomics* **67:** p256-67.

Stewart E. A., McKusick K. B., Aggarwal A., Bajorek E., Brady S., Chu A., Fang N., Hadley D., Harris M., Hussain S., Lee R., Maratukulam A., O'Connor K., Perkins S., Piercy M., Qin F., Reif T., Sanders C., She X., Sun W. L., Tabar P., Voyticky S., Cowles S., Fan J. B., Cox D. R., and et al. (1997). An STS-based radiation hybrid map of the human genome. *Genome Res* **7:** p422-33.

Tilford C. A., Kuroda-Kawaguchi T., Skaletsky H., Rozen S., Brown L. G., Rosenberg M., McPherson J. D., Wylie K., Sekhon M., Kucaba T. A., Waterston R. H., and Page D. C. (2001). A physical map of the human Y chromosome. *Nature* **409:** 943-945.

Vollrath D., Foote S., Hilton A., Brown L. G., Beer-Romero P., Bogan J. S., and Page D. C. (1992). The human Y chromosome: a 43-interval map based on naturally occurring deletions. *Science* **258:** p52-9.

Wang Y., Prade R., Griffith J., Timberlake W. E., and Arnold J. (1994). A fast random cost algorithm for physical mapping. *Proc. Natl. Acad. Sci. USA* **91:** 11094-11098.

Chapter 4

Prediction of radiation hybrid linkage likelihoods for multicopy markers

Charles Tilford, David Page

Introduction

In preceding chapters we have discussed the reasons why genomic maps typically use only single copy markers, and why it was necessary to ignore this restriction when constructing maps for the Y chromosome. For the YAC map, it was possible to localize multicopy markers due to the limited scope of each particular YAC. In these cases, it would be uncertain as to how many copies of the marker were present within a particular YAC, but at the same time it would be known with certainty that the marker was present in two or more locations due to their presence in clearly non-overlapping YACs.

Mapping multicopy markers in a radiation hybrid panel would pose new challenges. Localization of novel multicopy markers would hinge to a large extent on linkage to previously mapped loci. In the case of new subtraction-generated markers, knowledge of the location of the parent YAC could also be used to limit the initial scope of a marker. These observations could aid in the gross localization of an STS, but did not address the larger issue of calculating the precise linkage probability between a marker and potential neighbors. In particular, the algorithms available for calculating linkage probabilities assumes that both markers in the considered pair are either haploid or diploid. The Y chromosome harbors sequences that, as judged by their ability to be amplified by an STS primer pair, range from single copy in the genome up to 40+ copies in the case of TSPY (Tyler-Smith *et al.*, 1988). In addition to the wide range of copy numbers that individual

markers might have, the difference in copy number in a pairwise comparison could also be quite large.

This last consideration could be particularly troubling. The Y chromosome is a patchwork of regions that contain Y-specific single-copy, X/Y homologous and Y-specific repetitive regions. These disparate areas frequently adjoin one another - in such regions, it would be expected that linked markers would have dramatically different copy numbers. To illustrate the problem, consider hypothetical haploid marker A, present once on the chromosome. Right next to A (at an effective distance of zero) is marker B. However, marker B also has two other, unlinked copies present on the Y chromosome. The vectors for such a marker pair might appear as follows:

```
Marker A   0001000000010000000010000000001000000001000000000...
Marker B   0001001000010010001110010001001000100000100001001...
                 *       *   **   *   *      *          *  *
```

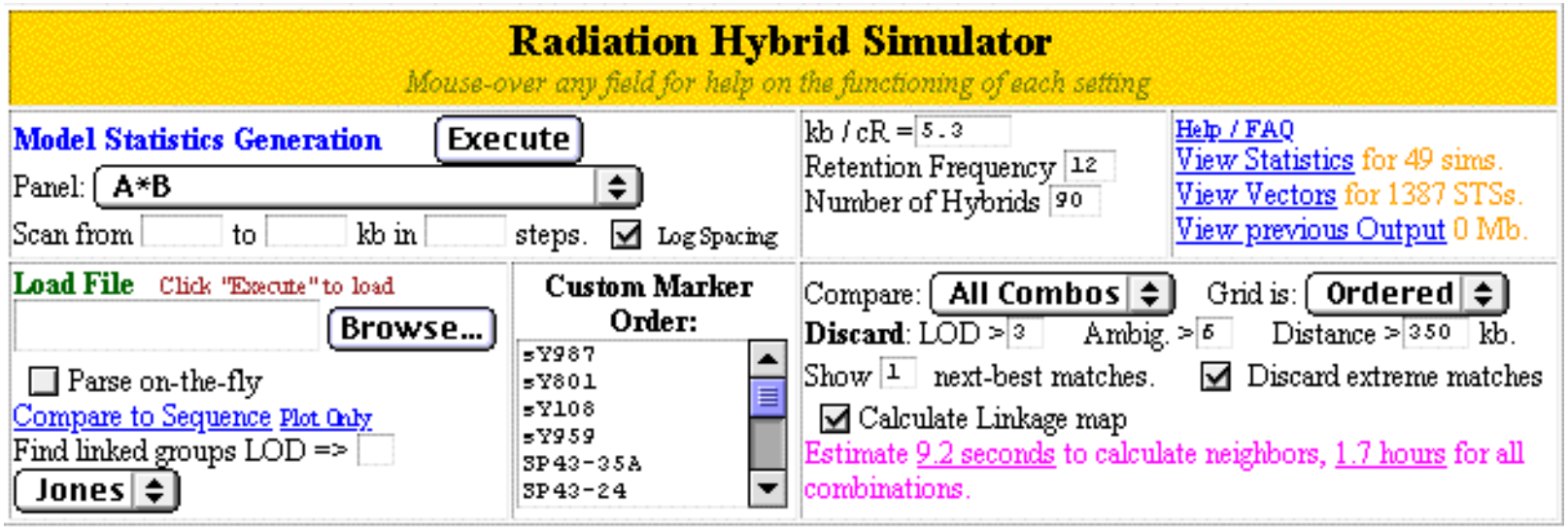
Every hybrid that is positive for A is also positive for B (since that tightly linked copy of B is always seen on any fragment containing A). However, there are 2 additional copies of B that are being randomly retained by hybrids. These copies generate additional positives in the B vector (noted by asterisks). At each of these positions, the hybrid pair shows discordance - A is negative, but B is positive. For both haploid and diploid algorithms, this discordance is interpreted as distance between the markers. With prior knowledge, we would describe A and B as being absolutely linked (with additional unlinked copies of B). Algorithmically, however, the two markers would be determined to have significant distance between them, and may even be indicated as being unlinked. However, consideration of the vector would lead an otherwise naive investigator to conclude that A and B are likely associated - after all, everywhere A is positive, B is as well.

To address this issue statistically, predictive software was designed to determine linkage between a pair of markers without making assumptions as to their copy number. This project began as software for simulating the behavior of multicopy markers in radiation hybrid panels. During the course of development, it was realized that the simulation provided predictive power as well. While the software now serves as a powerful tool for calculating linkage between radiation hybrid vectors, it is still referred to as "RH Simulator" based on this initial function.

Methods and Algorithms

Programming environment. The vast majority of the software is composed in a single file written in Perl 5. The software is accessed from a web browser, which serves as a front-end for providing user input and graphical output (Figure 1). Graphics are presented as GIF images generated using the GD library (Stein, 2000). The user specifies variables and support files (such as vector definitions) which are stored on the server in a temporary folder. Frequently-called, computationally intensive tasks, such as the actual calculation of probabilities, are coded in C and accessed from Perl using the XS interface.

Model construction. At the core of RH Simulator are models. A model serves as a description of the organization of a pair of markers (always referred to as A and B) in a genome. The model merely describes linkage, or lack thereof, and does not make any inferences as to the relative order of the markers. There are three possible components to a model: **A**, indicating a single, unlinked copy of A; **B**, similarly for marker B; and **A-B** indicating a linked pair of A and B. Note that self-linked components such as A-A or B-B are not considered. Also absent are components that specify linkage for more than two markers (e.g. A-B-A). The restriction to just three components was



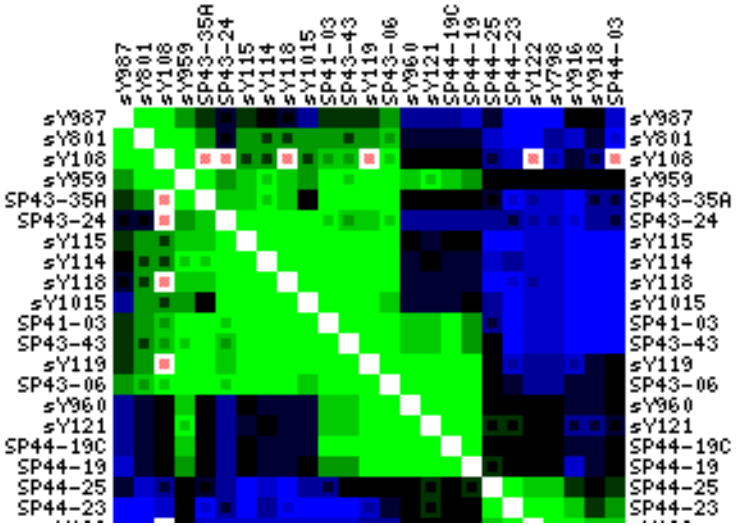


Figure 1. User interface for RH Simulator. A CGI-based web interface is used to interact with the algorithm. The user may specify external files to be loaded by the software, set constants such as the retention frequency, and define parameters for the simulation. Depending on settings chosen, results are presented as tables or graphical representations.

made first to limit the complexity of the algorithms and computations ultimately needed, and second because it was assumed (and later borne out by empirical evidence) that such fine variations in model composition would ultimately have little impact on calculated probability.

Components are combined to form the final models. Examples of models include: A-B (simple haploid linkage); A-B A-B (simple diploid linkage); A-B B B (haploid linked to triploid, as described in the introduction), etc. There are an infinite number of possible models, so only those models considering copy numbers of seven or less are considered. Also, if more than one A-B component is present, the distance between A and B is assumed to be the same for each component (again, a step taken to minimize the amount of computations).

Model probability curves. Any unambiguous vector comparison between markers (A and B) can be reduced to four values - the number of hybrid pairs that are both positive (AB), both negative (00), A positive but B negative (A0) and B positive and A negative (B0). These values ultimately reflect on the distance, or probability of linkage, between the two markers. Given a particular model, the probability that any one of these four classes will appear in a given hybrid pair can be calculated. To do so, it will be necessary to know the retention frequency (RF) of the panel, as well as the probability of X-ray separation between a pair of linked markers (θ). This theta value is dependent on two factors - the actual distance between the markers (d), and the number of kilobases per centiray (kb/cR), which is assumed to be constant for the panel and will be applied in the term λ:

$$\lambda = 0.01 / (kb/cR) \qquad [1]$$

This value is determined empirically. From analysis of the USP9Y region, a largely single-copy area of the Y chromosome which was sequenced

relatively early in the Y sequencing project, the TNG RH panel was determined to represent roughly 5.3 kb/cR. The theta value may then be calculated as a function of physical distance d (kb):

$$\theta = 1 - e^{-\lambda d} \quad [2]$$

The probabilities are then first calculated for the isolated components (A, B, or A-B) of the models. That is, if a model was composed of just a single component, then the probability of generating each of the four classes would be as follows:

For component A, the probabilities simply hinge on whether A is retained (at the retention frequency, RF) or lost (1-RF). It is impossible to generate classes containing B:

$$p00 = 1-RF \quad [3]$$

$$pA0 = RF \quad [4]$$

$$pB0 = pAB = 0 \quad [5]$$

For component B, the considerations are identical:

$$p00 = 1-RF \quad [6]$$

$$pB0 = RF \quad [7]$$

$$pA0 = pAB = 0 \quad [8]$$

For component A-B, the probability that a break might occur needs to be considered. If a break does occur (θ), the behavior of the two fragments must be accounted for, otherwise ($1-\theta$) the single fragment containing the both markers must be modeled:

$$pAB = (1-\theta)RF + \theta RF^2 \quad [9]$$

$$p00 = (1-\theta)(1-RF) + \theta(1-RF)^2 \quad [10]$$

$$pA0 = pB0 = \theta RF(1-RF) \quad [11]$$

These individual probabilities may then be combined to find the total probability for each class, in a given model at a given distance. In the

following calculations, products indicate the product of indicated probabilities for each component (the As, Bs and A-Bs) present in the model.

For a model to produce negative assays for both markers, each component must have been negative, so the total probability that the model produces a negative result is:

$$p00_{Total} = \prod p00 \quad [12]$$

For A to be positive but B negative, then *at least one* component must have provided an A0. No component could provide B0 or AB, but contributions of 00 are fine. Defining pA|0 as the probability of generating either A0 or 00:

$$pA|0 = pA0 + p00 \quad [13]$$

We simply calculate the product of this sum over all components. Included in this product is the probability that all components contributed nothing (00), so we need to subtract that probability out:

$$pA0_{Total} = \left(\prod pA|0 \right) - p00_{Total} \quad [14]$$

The same rationale is used to calculate the B0 overall probability:

$$pB|0 = pB0 + p00 \quad [15]$$

$$pB0_{Total} = \left(\prod pB|0 \right) - p00_{Total} \quad [16]$$

Directly calculating the total probability of AB would be fairly complex were it not for the fact that we have just calculated the previous three probabilities. Given that the four classes are comprehensive, the total probability must sum to 1:

$$pAB_{Total} = 1 - p00_{Total} - pA0_{Total} - pB0_{Total} \quad [17]$$

These total probabilities are calculated for each model over a discrete series of distances, from 0 up to a value *past* the maximum distance at which reliable linkage could be expected. This distance would correspond to a theta score of 0.6, which in the case of TNG (solving for equation 2) yields a

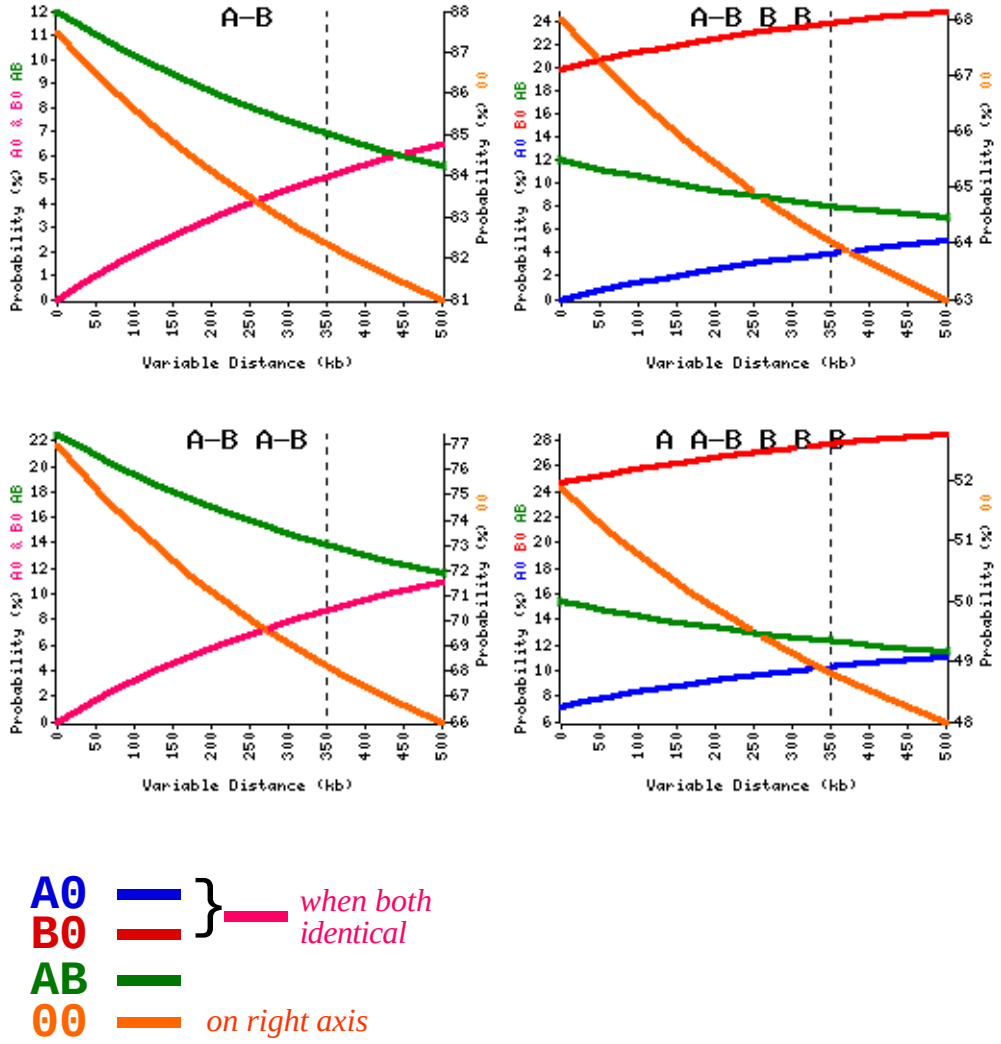


Figure 2. Probability curves for several models.
The probability of generating a hybrid of one of the four classes at various distances is shown for four models. Dashed line (350kb) indicates presumed distance past which linkage may not be ascertained. Note that the 00 class (orange) is plotted on the right axis, and that the A0 and B0 classes are plotted in pink when they yield identical probabilities. (RF = 0.12, 5.3 kb per cR)

distance of roughly 350kb. In the analysis of the TNG data a logarithmically spaced series of 12 values ranging from 0-500kb was used. Figure 2 shows examples of probability curves for several diverse models.

Calculation of model probability. The probability curves may now be used to calculate the probability that a given model would produce the observed experimental data. First, a vector pair is tallied to determine the representation of each of the four classes such that "a" represents the number of hybrid pairs in the A0 class, "b" the number that are B0, "ab" those that are AB and "x" those that are 00. The total number of hybrids present in the panel is then represented by "N," where:

$$N = a + b + ab + x \qquad [18]$$

The probability of observing a A0 classes is then simply the overall A0 probability raised to the power a. However, there are many different ways for those hybrids to be distributed in a vector, and this must be accounted for by factoring in the number of discrete possibilities as "N choose a" or defined more generically as "n choose k":

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad [19]$$

As we factor in the remaining probabilities, the number of "free" hybrids has been diminished by previous choices, and this must be factored into the calculation of the number of probabilities. The overall probability of observing the data from a given model is then:

$$P = pA0_{Tot}^{a}\binom{N}{a}pB0_{Tot}^{b}\binom{N-a}{b}pAB_{Tot}^{ab}\binom{N-a-b}{ab}p00_{Tot}^{x}\binom{N-a-b-ab}{x} \qquad [20]$$

Since we are ultimately "choosing" all possible hybrids, the four representations of equation 19 shown above ultimately collapse into the multinomial coefficient:

$$\binom{N}{a}\binom{N-a}{b}\binom{N-a-b}{ab}\binom{N-a-b-ab}{x} = \binom{N}{a\ b\ ab\ x} = \frac{N!}{a!b!ab!x!} \qquad [21]$$

Substitution of 21 into 20 then produces the most compact form of the equation:

$$P = \frac{pA0^{a}_{Tot}\, pB0^{b}_{Tot}\, pAB^{ab}_{Tot}\, p00^{x}_{Tot}\, N!}{a!b!ab!x!} \qquad [22]$$

Equation 22 can now be used to calculate the probabilities that a given model would produce the observed data at each of the specifieddistances. The highest of these probabilities is then reported for the model, along with the distance at which that optimal probability was found. If the distance exceeds the maximum likely distance for linkage, the model is not reported at all. This is because at extreme distances, the model begins to statistically behave as a totally unlinked model (lack of linkage typically being due to large distance between markers). Because RH simulator explicitly considers unlinked models (that is, models lacking any A-B components), inclusion of linked models at extreme distance would duplicate the values reported for unlinked sets, and may lead the user to assume linkage where none exists.

Data provided to the algorithm often contains ambiguous results. These hybrids, scored as '2', were positive in one assay and negative in another; a 2 in the vector could thus be interpreted as *either* a '0' or a '1'. In considering a vector pair with a total of "t" ambiguities in the two vectors, there are then 2^t possible unambiguous vector pairs. The software determines each of these potential pairs and reduces them to the appropriate set of four pairwise outcomes. Some of these outcome sets will be the same (that is, even if the vector pairs are different, the number of AB, 00, A0 and B0 classes will be identical), so these are identified and grouped to minimize the following computational steps. The probabilities and distances at which they occur are calculated for each of these sets and averaged (with appropriate weightings being applied for sets that occurred more than once), and the average probability and distance, with standard deviations, are reported. In this

manner ambiguity will result in an increased uncertainty in probability / distance calculations, as indicated by higher standard deviations.

Identification of STS sites on sequence. "Electronic PCR" (ePCR) was performed on known Y chromosome sequence to identify likely STS sites. For a contiguous stretch of sequence, all potential primer binding sites for an STS were determined in both frames. Primers were allowed to have one mismatch in any position. "Forward" primer sites were then compared to the "reverse" site. Provided that the primers were properly oriented and within 50-1000bp, the location was tagged as a site that would produce a band on STS amplification.

Results and Discussion

During the development of this software, it was hoped that the calculations would not only demonstrate whether two markers were linked, but also which model represented their likely organization on the chromosome. In other words, ideally the most probable model would be a reliable description of both the copy number and linkage state of a pair of markers. Unfortunately, this proved not to be the case - the best model had a probability that was typically very similar to the next best models (often within LOD 1-2). Table 1 shows example output from the simulator, with six neighboring marker pairs being computed. The models are sorted by probability, with output truncated once 3 logs have been displayed from the best model. It can be seen that there are typically multiple models within LOD 3 of the best model.

The similarity in certainty between models is more than likely due to an inability to extract copy number from a single vector. STSs with more copies are expected to generate more positive results in a vector. Given a retention

frequency of RF, the probability that an individual hybrid tested with a marker with c copies will be negative (P_{0c}) or positive (P_{1c}) is:

$$P_{0c} = (1-RF)^c \quad [22]$$

$$P_{1c} = 1 - P_0 \quad [23]$$

The overall probability (P_c) that a marker with c copies would produce a vector with x positives (out of a total of N hybrids) would then be expressed as:

$$P_c = P_{0c}^{N-x} P_{1c}^{x} \binom{N}{x} \quad [24]$$

If we apply this equation to the TNG panel, with a retention frequency of 12% and 90 hybrids, we find that there is significant overlap in probabilities. Figure 3 shows the probability curves for observing a particular number of positive hybrids for marker copy numbers between 1 and 10. Probabilities have been expressed as base 10 logarithms. It can be observed that while in most cases there is a single, most probable copy number, there are typically 3-4 additional copy numbers with similar probability. For example, consider the region around 20 positive hybrids (highlighted on the figure). The most probable copy number is 2, with a 10^{-1} probability of generating 20 positives. However, 3 copies have a probability of $10^{-1.9}$, single copy of $10^{-2.6}$, and even four copies should still be considered likely at $10^{-3.8}$ if we wish to have LOD 3 certainty in our estimates.

Copy number is the single greatest determinant affecting the probability of a model generating observed data, given that it has a direct and significant impact on the probability of observing a positive, which lies at the core of pairwise comparisons. Because copy number prediction is uncertain, so is the assignment of a single, specific model. In fact, the uncertainty in model assignment tracks the uncertainty of the copy numbers for the two markers very closely. Computing the copy number probability for the four

Marker Pair	Model	log Probabil at	Distance kb
sY1 *vs.* sY2	A-B B	1.95 ± 0.12 at	0 ± 0 kb
	A-B B B	3.04 ± 0.16 at	0 ± 0 kb
	A-B B B B	5.44 ± 0.30 at	0 ± 0 kb
sY2 *vs.* SP01-25A	A A-B B	2.86 ± 0.10 at	19 ± 21 kb
	A A-B B B	3.16 ± 0.14 at	0 ± 0 kb
	A A A-B B B	4.64 ± 0.24 at	0 ± 0 kb
	B B-A A A	4.64 ± 0.23 at	0 ± 0 kb
	A A-B B B B	4.88 ± 0.28 at	0 ± 0 kb
	A-B B B	4.93 ± 0.26 at	286 ± 25 kb
	A A A-B B B B	6.15 ± 0.34 at	0 ± 0 kb
SP01-25A *vs.* SP01-25B	A-B A-B	2.61 ± 0.08 at	204 ± 24 kb
	A-B A-B A-B	3.06 ± 0.06 at	172 ± 15 kb
	A A-B B	3.44 ± 0.16 at	0 ± 0 kb
	A-B A-B A-B A-B	4.82 ± 0.12 at	145 ± 18 kb
	B B-A A A	4.92 ± 0.35 at	0 ± 0 kb
	A A-B B B	5.25 ± 0.38 at	0 ± 0 kb
	B-A A	5.30 ± 0.35 at	157 ± 19 kb
	A-B	5.56 ± 0.32 at	250 ± 26 kb
	A-B B	6.02 ± 0.39 at	175 ± 27 kb
SP01-25B *vs.* sY3	A-B A-B	1.97 ± 0.09 at	38 ± 14 kb
	A-B A-B A-B	2.49 ± 0.15 at	30 ± 0 kb
	A-B A-B A-B A-B	4.15 ± 0.17 at	30 ± 0 kb
	A-B	4.29 ± 0.02 at	38 ± 14 kb
	A-B A-B A-B A-B A-B	6.45 ± 0.19 at	30 ± 0 kb
sY3 *vs.* sY5	A-B A-B	2.48 ± 0.05 at	162 ± 22 kb
	A-B A-B A-B	3.16 ± 0.15 at	143 ± 21 kb
	A A-B B	3.85 ± 0.33 at	0 ± 0 kb
	A-B	4.95 ± 0.19 at	202 ± 29 kb
	A-B A-B A-B A-B	5.06 ± 0.24 at	122 ± 17 kb
	B-A A	5.11 ± 0.53 at	122 ± 21 kb
	B B-A A A	5.66 ± 0.47 at	0 ± 0 kb
sY5 *vs.* sY6	B-A A	1.93 ± 0.19 at	16 ± 27 kb
	B-A A A	2.99 ± 0.42 at	14 ± 24 kb
	A-B A-B	5.02 ± 0.47 at	231 ± 36 kb

Table 1: Sample output from RH simulator. Pairwise comparisons between 7 neighboring markers are shown. The simulator has calculated the most likely models, and displayed their probability ($-\log_{10}$) as well as the distance at which that probability occured.

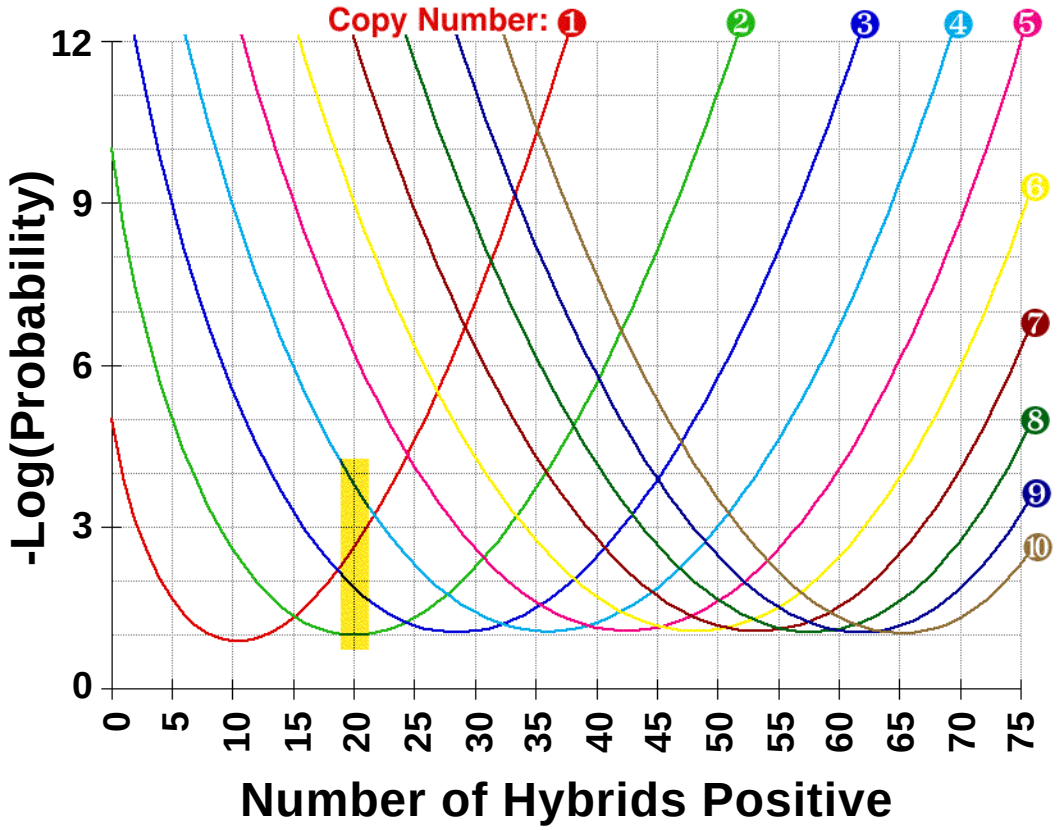


Figure 3. Probability of observing a particular number of positive hybrids for markers with various copy numbers. Probabilites are expressed as negative logarithms. Each curve represents the probabilities for a specific copy number, indicated at the right terminus of each curve. The highlighted region around 20 positive hybrids is used as an example in the text.

	Probability ($-\log_{10}$) to derive observed data from a copy number of:							
Marker	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
sY1	**1.03**	**3.39**	**6.74**	**10.50**	**14.48**	**18.59**	**22.80**	**27.08**
sY2	**1.99**	**1.06**	**2.36**	**4.57**	**7.27**	**10.29**	**13.51**	**16.88**
SP01-25A	**3.70**	**1.12**	**1.39**	**2.83**	**4.89**	**7.36**	**10.09**	**13.01**
SP01-25B	**3.31**	**1.05**	**1.53**	**3.12**	**5.32**	**7.89**	**10.72**	**13.73**
sY3	**3.31**	**1.05**	**1.53**	**3.12**	**5.32**	**7.89**	**10.72**	**13.73**
sY5	**2.94**	**1.01**	**1.70**	**3.44**	**5.77**	**8.45**	**11.38**	**14.48**
sY6	**0.90**	**2.27**	**5.00**	**8.30**	**11.89**	**15.68**	**19.59**	**23.60**

Table 2: Copy number likelihoods for selected markers. The seven markers presented in Table 1 are analyzed. the probability ($-\log_{10}$) of observing the data given a particular copy number (between 1 and 8) is presented for each marker. Shaded cells represent values within 3 logs of the most likely probability.

markers shown in the example above results in Table 2. Shaded cells indicate probabilities within LOD 3 of the most probable copy number. Comparison with Table 1 reveals that the uncertainty in model likelihoods almost completely overlaps the uncertainty in copy number likelihoods.

This discovery was a disappointment, as it was hoped that the simulator would be able to provide more specific information on the nature of the markers. However, the simulator has still proven useful in calculating the probability of linkage between a pair of markers. There are two general types of models: linked, which assume at least some linkage between the two markers; and unlinked, which assume that all markers are totally unlinked. Instead of comparing the best model to all other models, the overall best model is compared to the best model of the other linkage class. In this case the difference in probabilities is often very significant, such that a pair of markers can be said with high confidence (LOD 4+) to be linked or unlinked.

Figure 4 shows the results from such an analysis on part of the TNG Y dataset. Each pairwise comparison indicates the likelihood of linkage as a color value, with green values indicating high probability of linkage and blue values a high probability that the markers are unlinked. The values are determined by subtracting from the probability of the best model the probability of the best model of other linkage class (i.e. either linked minus unlinked, or unlinked minus linked). Darker values indicate lower probabilities, with black representing an equal likelihood that the markers are linked or unlinked. Uncertainty is indicated by a smaller square in the center of each data point. This value is calculated by including the standard deviation in the calculation (e.g. (unlinked-SD)-(linked+SD)), and always represents a probability equal to or less likely than that calculated from the average probabilities.

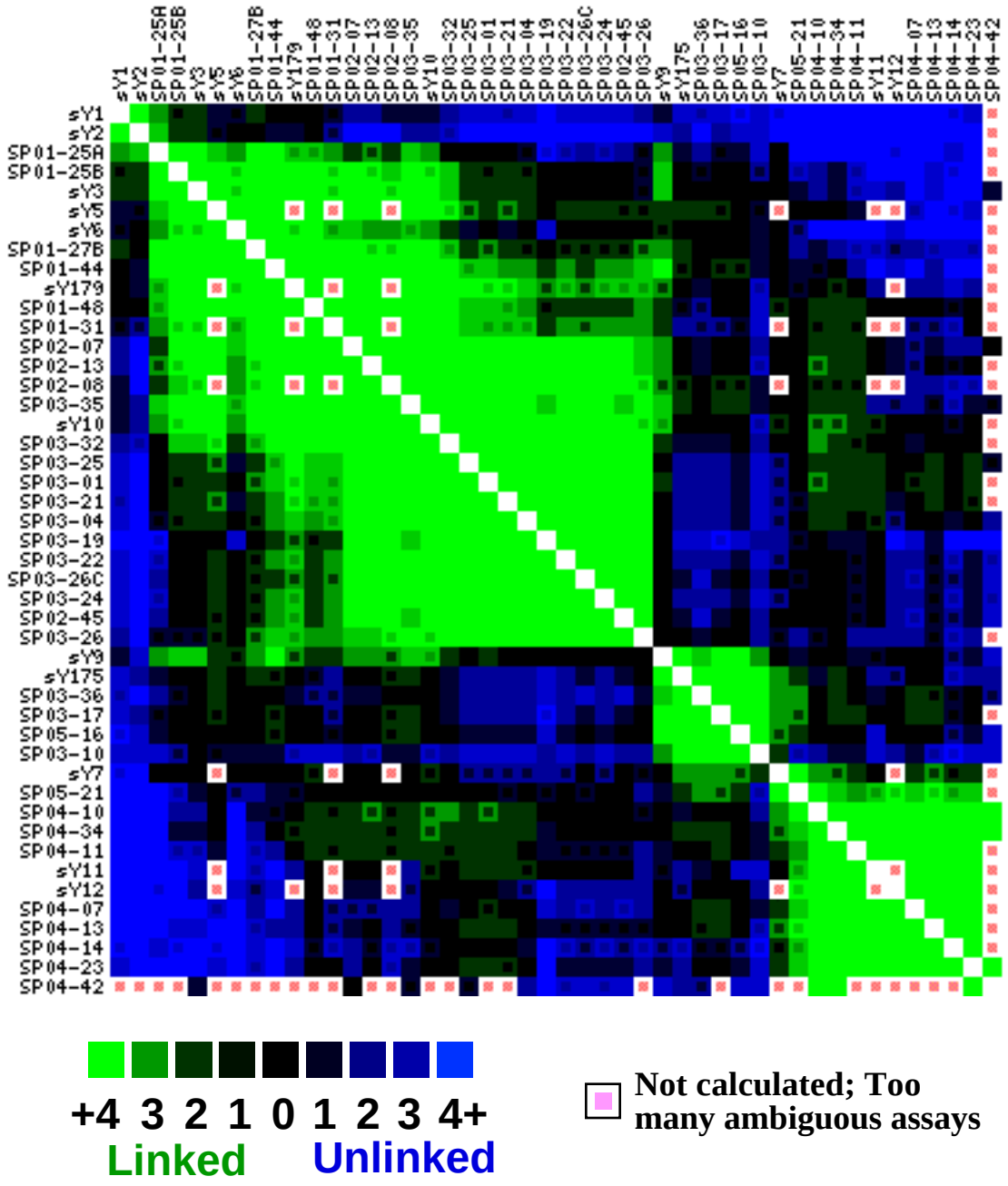


Figure 4. Graphical representation of linkage likelihood for a subset of the Y chromosome. Each pair of markers has a likelihood of being linked or unlinked, which is represented as a colored square. The likelihood is calculated as the difference between the probabilities of the most likely linked model and the most likely unlinked model. In cases where there is significant standard deviation in the calculation of the model probabilities, the lowest likelihood between the probabilities (as adjusted by the standard deviation) is also calculated, and is represented by the smaller internal square (a value that is always less significant than the difference of the averages).

As can be seen, linked clusters of markers are clearly indicated as solid blocks of green. The figure is symmetrical about the diagonal, which has been obscured (although the diagonal should be solid green to represent absolute linkage, this is non-informative and ultimately distracting). Because the computational time for each pair is proportional to the square of the number of ambiguous hybrids, vector pairs exceeding a user-specified limit of ambiguous assays (here 6) are not analyzed, and are shown as a white square with pink dot.

Recalling that the impetus for developing this software was to detect linkage between high-copy markers that might be over-looked by standard algorithms, it was then hoped that the simulator would be able to predict higher probabilities of linkage for high copy-number markers. However, when the likelihood calculations are compared to those derived using the functions from Jones (Jones, 1996), the opposite was noted. Figure 5 shows the result of such a comparison for likelihood calculations in 550 marker pairs (selected for above average PCR assay quality, in order to avoid artifacts due to poor data quality). The two algorithms show the same trend, which is reassuring (Jones calculates only the probability of linkage, so the minimum score possible is 0. The simulator considers the probability that the markers are unlinked as well. These values are indicated as negative scores, hence the "squashed tail" on the left of the plot). However, when they differ, the Jones score is almost always higher (more probable) than that produced by the simulator - most of the data points fall outside the green area (which denotes points where the Jones value is lower than the simulator value).

By providing the simulator with many models of differing copy number, we have not only given it more opportunity to find linkage, but also the option of choosing more models without linkage. While the linkage

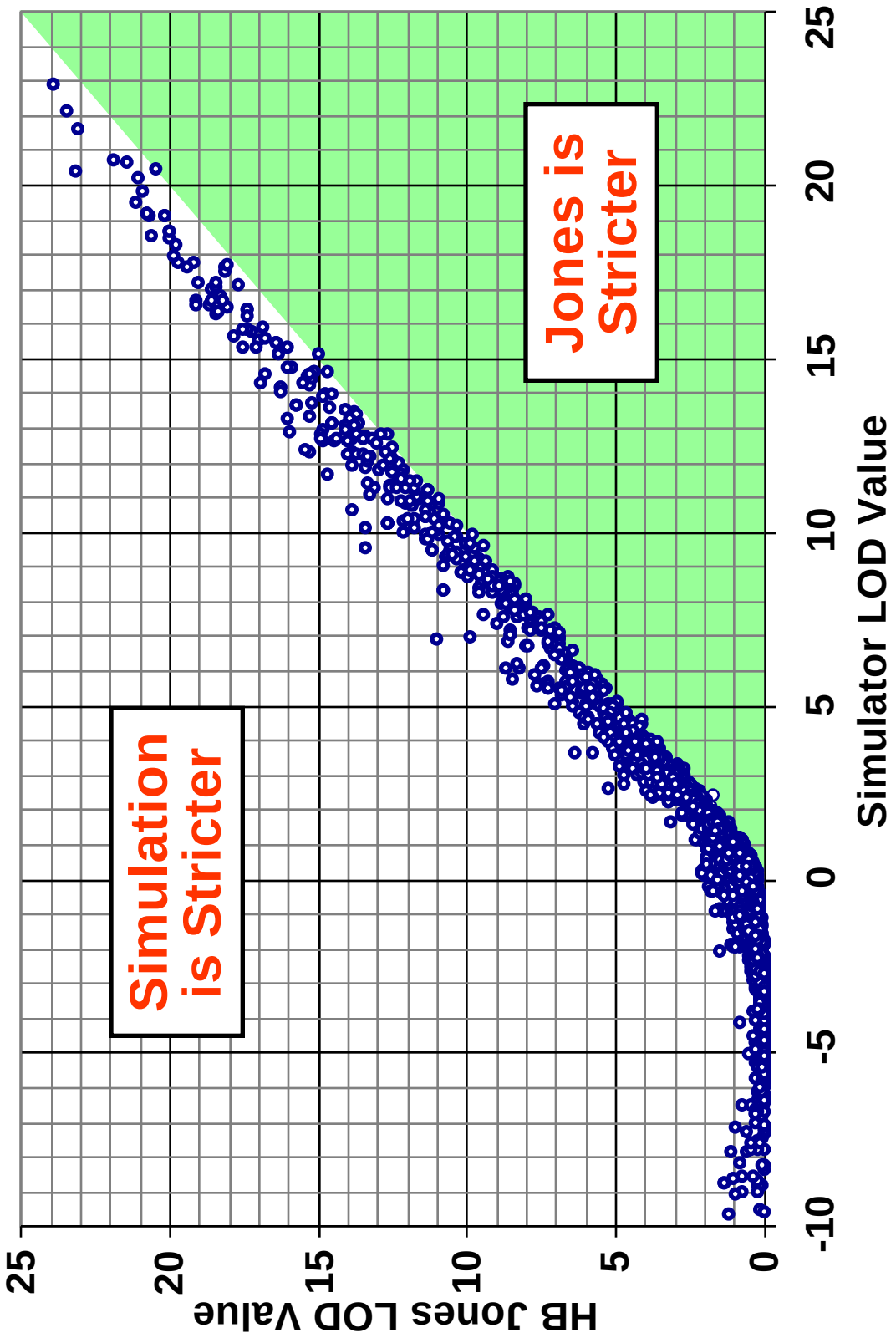


Figure 5. Comparison of LOD scores generated by the RH Simulator and methods described by HB Jones

likelihood for multicopy markers will be higher than standard methods, there are multiple null-hypothesis (unlinked) models that will also be ranked with higher probability. Because the final LOD score represents not simply the probability of the linked model, but the ratio of it with the best unlinked model, this overall LOD is lower. Again, this is a disappointment, but likely represents a more accurate measure of the linkage probabilities between markers than resorting to algorithms that assume only haploid or diploid marker content.

Sequencing of the Y chromosome is essentially complete, with only a few small gaps remaining to be closed (Kuroda-Kawaguchi *et al.*, 2001b). It is then possible to compare the linkage likelihoods generated by the simulator to the actual location of marker pairs in the sequence. Only markers of reasonable quality (as scored during PCR typing) were used in this comparison to minimize errors due to poor STS quality. For marker pairs present on a contiguous segment of DNA (as determined by ePCR), the distance between the closest two members was plotted against the linkage likelihood. The results are shown in Figure 6, with distance plotted on a logarithmic scale.

The trend is as expected, with increasing likelihood of linkage corresponding to decreased distance between markers. There are occasionally marker pairs that are in very close proximity, but have a low likelihood of linkage. These pairs have been analyzed in an attempt to find a consistent feature that might explain their low linkage scores, but without success. It is possible that they represent false ePCR results - while the electronic conditions for a functional STS were met at that point in the sequence, *in vitro* the site fails to serve as a PCR target. It is also possible that the difference is due to sequence differences between the male represented by the

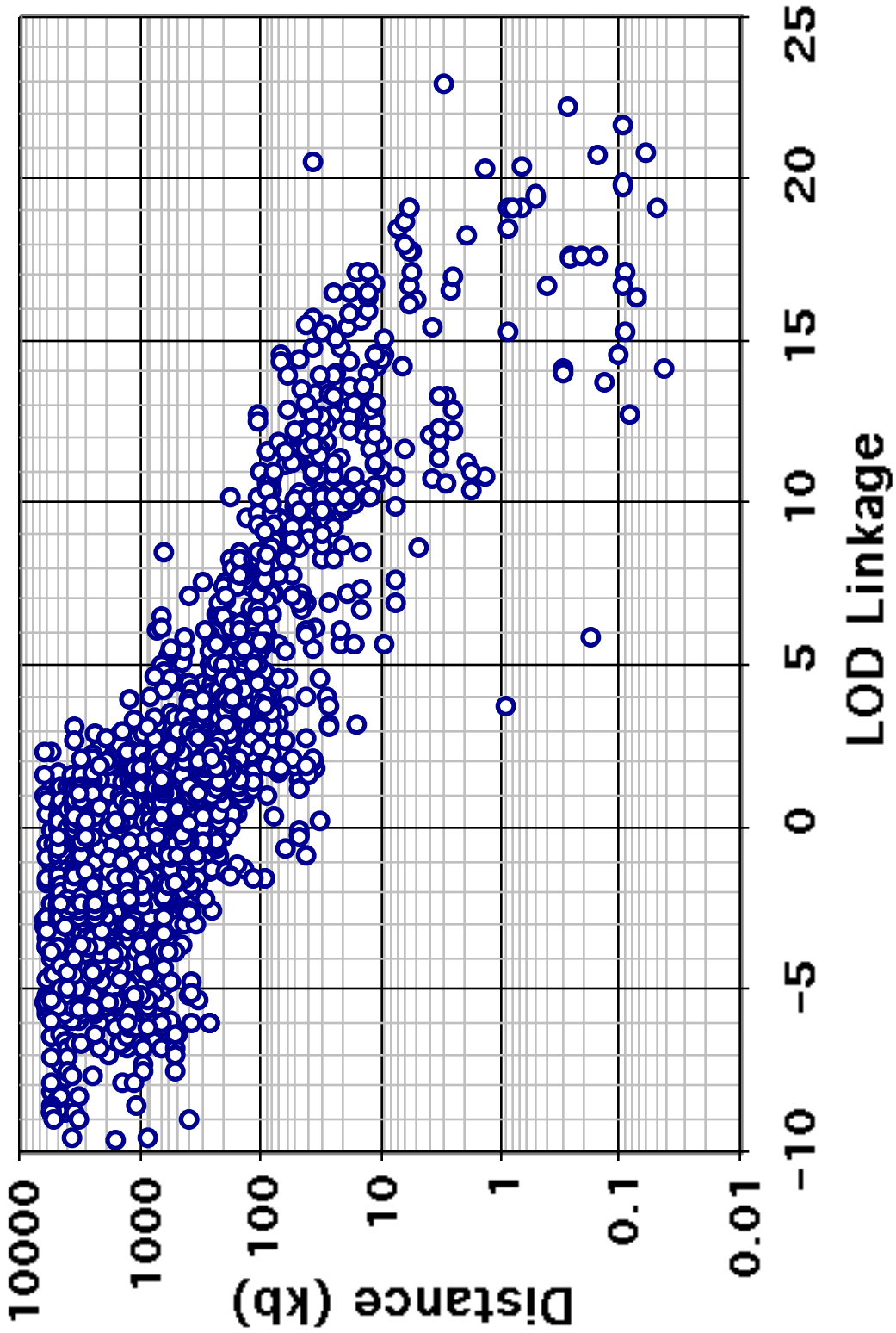


Figure 6. Comparison of Simulator linkage likelihood with physical distance, as determined by ePCR.

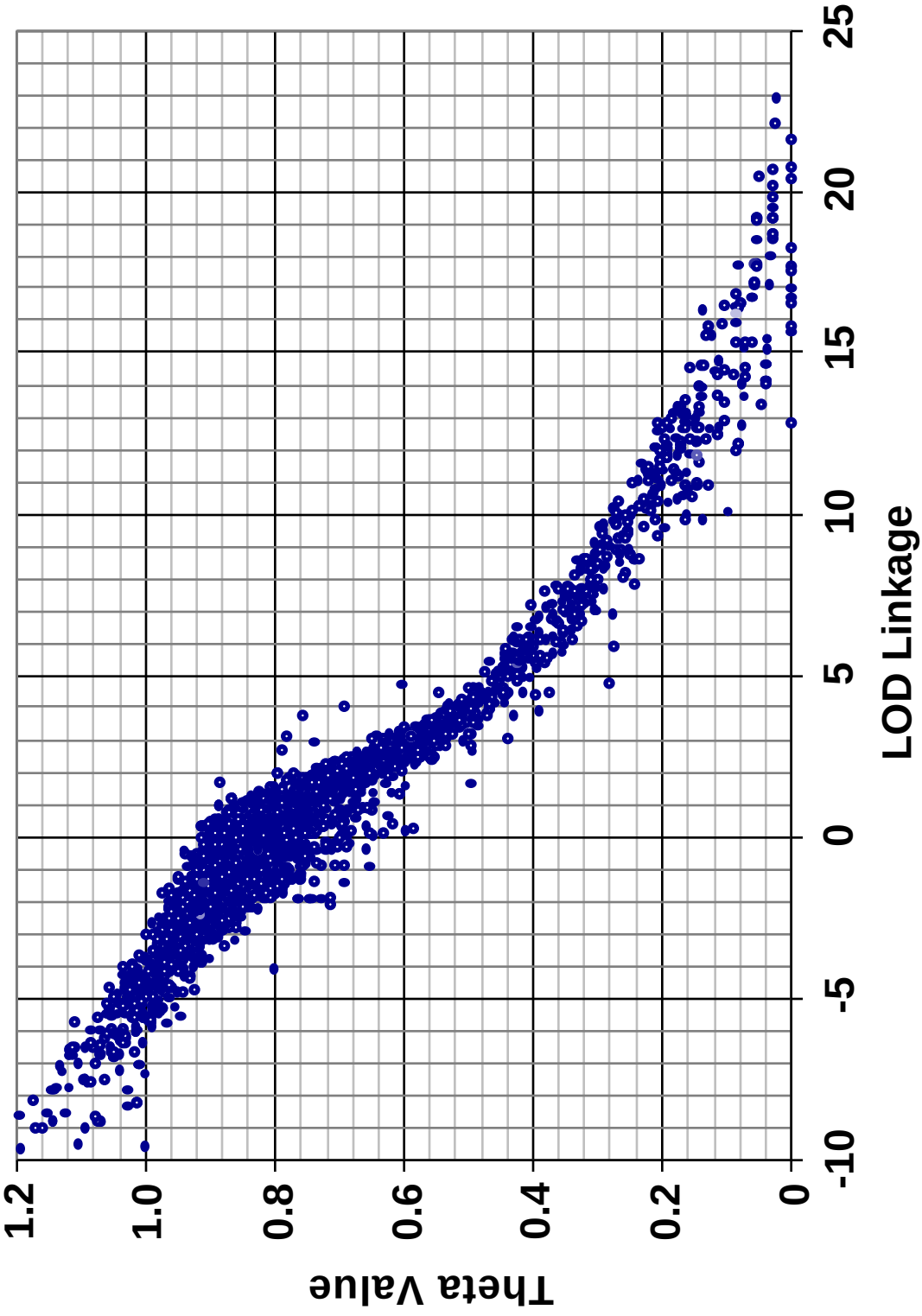


Figure 7. Comparison of Simulator linkage likelihoods with theta values.

BAC library, and the one used as a genome donor for the TNG panel. Another, intriguing possibility is the presence of X-ray hot spots in the radiation hybrid panel. It is possible that some areas have greatly elevated likelihood of suffering X-ray induced breakage. Markers flanking such regions would be separated at a much higher frequency than expected for their distance, and would be predicted to have an artificially low likelihood of linkage.

As a final comparison, the LOD scores generated by the simulator were compared to the theta values calculated for the same marker pairs. As can be seen in Figure 7, the two values track very closely. Theta scores represent a reasonable predictor of linkage likelihood (at least for this subset of the data, which represents markers of average or above quality), although the relationship is not linear. In particular, the vast majority of marker pairs with theta scores exceeding 0.6 have linkage likelihoods below LOD 3.

In summary, the simulator represents an accurate method for calculating linkage between markers, as judged by comparison to actual sequence, and by agreement with standard haploid algorithms. It does not appear to represent a significant improvement over algorithms that assume a haploid genome, however.

References

Jones H. B. (1996). Pairwise analysis of radiation hybrid mapping data. *Ann Hum Genet* **60:** p351-7.

Kuroda-Kawaguchi T., Skaletsky H., Minx P., Brown L., Rozen S., Wilson R., Waterston R., and Page D. (2001). The DNA sequence of the human Y chromosome. *Unpublished.*

Stein L. D. (2000). GD.pm - Interface to Gd Graphics Library, http://stein.cshl.org/WWW/software/GD/.

Tyler-Smith C., Taylor L., and Muller U. (1988). Structure of a hypervariable tandemly repeated DNA sequence on the short arm of the human Y chromosome. *J Mol Biol* **203:** 837-48.

Chapter 5

A physical map of the human Y chromosome

Charles A. Tilford*, Tomoko Kuroda-Kawaguchi*, Helen Skaletsky*, Steve Rozen*, Laura G. Brown*, Michael Rosenberg*, John D. McPherson†, Kristine Wylie†, Mandeep Sekhon†, Tamara A. Kucaba†, Robert H. Waterston† & David C. Page*

** Howard Hughes Medical Institute, Whitehead Institute, and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA*

† Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, 4444 Forest Park Boulevard, St. Louis, Missouri 63108, USA

The non-recombining region of the human Y chromosome (NRY), which comprises 95% of the chromosome, does not undergo sexual recombination and is present only in males. An understanding of its biological functions has begun to emerge from DNA studies of individuals with partial Y chromosomes, coupled with molecular characterization of genes implicated in gonadal sex reversal, Turner syndrome, graft rejection and spermatogenic failure[1,2]. But mapping strategies applied successfully elsewhere in the genome have faltered in the NRY, where there is no meiotic recombination map and intrachromosomal repetitive sequences are abundant[3]. Here we report a high-resolution physical map of the euchromatic, centromeric and heterochromatic regions of the NRY and its construction by unusual methods, including genomic clone subtraction[4] and dissection of sequence family variants[5]. Of the map's 758 DNA markers, 136 have multiple locations in the NRY, reflecting its unusually repetitive sequence composition. The markers anchor 1,038 bacterial artificial chromosome clones, 199 of which form a tiling path for sequencing.

A low-resolution physical map of the Y chromosome was previously assembled by testing naturally occurring deletions and yeast artificial chromosome (YAC) clones for the presence or absence of 182 Y-chromosomal sequence-tagged sites (STSs)[3,6]. These STS markers were generated from Y DNA sequences selected at random, which promoted representative sampling of the entire chromosome[6]. Nonetheless, most randomly selected sequences proved unusable as map landmarks because they corresponded either to interspersed repetitive elements found throughout the genome or to male-specific repetitive sequences dispersed to many locations in the NRY.

To construct a high-resolution map, we generated additional STSs in a directed manner. To enrich for the single-copy sequences most useful as map landmarks, we systematically applied genomic clone subtraction, whereby a 'tracer' clone's DNA is depleted of sequences shared with a set of 'driver' clones[4]. We identified a tiling path of 57 YACs that collectively spanned the euchromatic NRY[3], and then carried out, in parallel, 57 subtractions, each employing one YAC as tracer and the remaining YACs (minus those overlapping the tracer YAC) as drivers. We sequenced a random sample of products from each of the 57 subtractions, identifying 308 additional STSs that proved useful in map assembly.

We used radiation hybrid mapping to integrate and order the random and subtraction-derived STSs. Large-fragment radiation hybrid panels offering long-range connectivity have been used to assemble human genome maps at 500–1,000-kilobase (kb) resolution[7–10]. To obtain greater resolution in the NRY, where the number of STSs appeared sufficient to cover the euchromatic region at an average spacing of 50 kb, we used a small-fragment radiation hybrid panel that had been used to build detailed maps of limited autosomal segments[11]. We tested this panel for all random and subtraction-derived NRY STSs. Nascent radiation hybrid linkage groups were ordered and oriented with respect to the centromere by positioning selected STSs on the existing map of natural deletions[6]. Additional STSs generated in our project's later phases were also tested. Ultimately, 513 STSs were positioned, at a resolution of around 50 kb, on a radiation hybrid map encompassing nearly the entire euchromatic NRY.

To prepare for sequencing the NRY, we used the radiation hybrid map as a scaffold for assembling contigs of bacterial artificial chromosome (BAC) clones. We screened a BAC library of human male genomic DNA with hybridization probes derived from NRY STSs. Through subsequent polymerase chain reaction tests of STS content, we assembled 1,038 BACs into contigs that, except for four small gaps, represented the whole NRY (see Fig. 1 in Supplementary Information). Many portions of this BAC map could be assembled only after the sequence of selected BACs had been determined and compared. Also, many NRY genes and extragenic sequences are known to have closely related counterparts on the X chromosome[12]. In many cases, it was initially unclear whether BACs identified using X-homologous NRY STSs, especially those from a 4-Mb region of 99% X–Y identity[13,14], derived from the NRY or from the X chromosome. We resolved these ambiguities by resequencing STSs from the BACs in question and comparing them to X- or Y-derived reference sequences. The resulting map of overlapping BACs and ordered STSs (Fig. 1 in Supplementary Information)

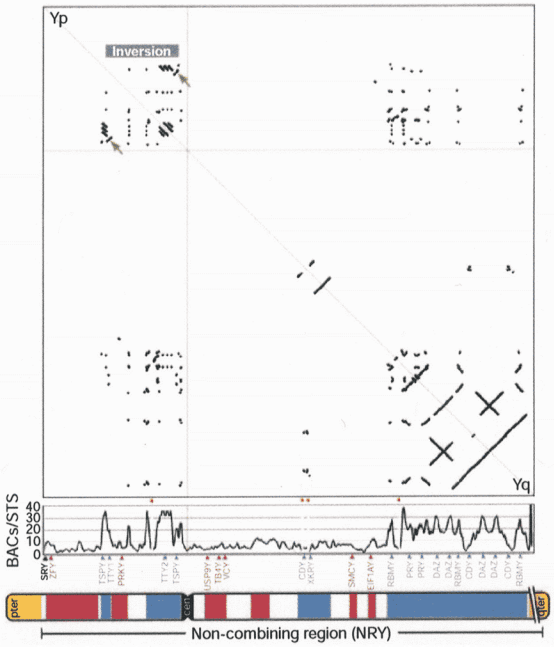


Figure 1 Repetitive structure of euchromatic NRY. Bottom, schematic of the Y chromosome, comprising large NRY flanked by pseudoautosomal regions (yellow). NRY is divided into euchromatic and heterochromatic (tan, shown truncated) portions, roughly 24 and 30 Mb, respectively. pter, short-arm telomere; cen, centromere; qter, long-arm telomere. Within euchromatic NRY, regions rich in NRY-specific amplicons (blue) or sequence similarity to X chromosome (red) are shown. Above chromosome schematic are positions of some NRY genes; most are found in amplicons (blue) or have X-linked homologues (red). Above genes is a plot of the average number of NRY BACs that contain each of the 758 STSs mapped (136 of these STSs at two or more locations) along euchromatic NRY. As expected, STSs in amplicon regions tended to be present in more BACs than STSs in X-homologous or unshaded regions. (Plotted values are local averages within sliding window of five consecutive STSs; values reflect all NRY BACs containing those STSs, not just BACs assigned to site indicated.) Some amplicon regions were under-represented in the BAC library; four gaps remain (red diamonds; ≤100 kb each) in BAC coverage of NRY. Top, STS-based dot plot of euchromatic NRY. Each dot reflects occurrence of a particular STS at two points in map (complete map shown in Fig. 1 in Supplementary Information). Dots fall almost exclusively within amplicon regions. Many repeats of entire groups of STSs are apparent, with lines parallel to light grey diagonal indicating direct repeats and lines perpendicular to light grey diagonal indicating inverted repeats. Green arrows, inverted repeats flanking 3.5-Mb inversion (see text). Pale red lines, centromere.

was extensively cross-checked against the radiation hybrid map and was further reinforced by restriction fingerprinting of all mapped BACs[15].

The greatest challenges were posed by massive, NRY-specific amplified regions (or amplicons), which comprise about one-third of the euchromatic NRY. Of the 758 STSs on which the map is built, 136 are present at two or more locations in the NRY. Although we avoided such repetitive STSs in favour of single-copy STSs wherever possible, substantial portions of the euchromatic NRY contained little or no single-copy sequence. For many such amplicons, BACs derived from different copies could not be distinguished by STS content or restriction fingerprinting[15]. In many cases, we distinguished among amplicon copies (and the BACs corresponding to them) by typing 'sequence family variants' (SFVs)[5]. SFVs are subtle differences (for example, single-nucleotide substitutions or dinucleotide repeat length alterations) between closely related but non-allelic sequences. We were analysing BACs from only one male's Y chromosome, so these subtle sequence differences could not represent allelic variants. In general, we identified SFVs only after comparing the DNA sequences of BACs that originated from distinct copies, despite having similar STS content. Thus, mapping and sequencing were inseparable, iterative activities in amplicon-rich regions.

The euchromatic NRY amplicons are diverse in composition, size, copy number and orientation (Fig. 1), with some occurring as tandem repeats, others as inverted repeats, and still others dispersed throughout both arms of the chromosome. The euchromatic amplicons are well populated with testis-specific gene families that may be critical in spermatogenesis (see Fig. 1 in Supplementary Information)[1,2].

One pair of amplicons is of particular interest in the context of human variation. Highlighted in Fig. 1 (arrows) are two units, each 300 kb long, that exist in opposite orientations on the short arm. These inverted repeats bound a region of around 3.5 Mb that occurs in one orientation (Fig. 1 in Supplementary Information) in the male from whom the BAC library was constructed, but in the opposite orientation in the existing map of naturally occurring deletions[6]. This may reflect variation among men for a 3.5-Mb inversion[1,16], perhaps the result of homologous recombination between the 300-kb inverted repeats flanking the inverted segment. Large Y-chromosome inversions are postulated to have been crucial in the evolution of the human sex chromosomes[12], and this 3.5-Mb inversion may be one of many massive NRY variants that exist in modern populations. □

Received 16 November; accepted 21 December 2000.

1. Vogt, P. H. *et al.* Report of the Third International Workshop on Y Chromosome Mapping 1997. *Cytogenet. Cell Genet.* **79,** 1–20 (1997).
2. Lahn, B. T. & Page, D. C. Functional coherence of the human Y chromosome. *Science* **278,** 675–680 (1997).
3. Foote, S., Vollrath, D., Hilton, A. & Page, D. C. The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. *Science* **258,** 60–66 (1992).
4. Reijo, R. *et al.* Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. *Nature Genet.* **10,** 383–393 (1995).
5. Saxena, R. *et al.* Four *DAZ* genes in two clusters found in the *AZFc* region of the human Y chromosome. *Genomics* **67,** 256–267 (2000).
6. Vollrath, D. *et al.* The human Y chromosome: A 43-interval map based on naturally occurring deletions. *Science* **258,** 52–59 (1992).
7. Hudson, T. J. *et al.* An STS-based map of the human genome. *Science* **270,** 1945–1954 (1995).
8. Gyapay, G. *et al.* A radiation hybrid map of the human genome. *Hum. Mol. Genet.* **5,** 339–346 (1996).
9. Stewart, E. A. *et al.* An STS-based radiation hybrid map of the human genome. *Genome Res.* **7,** 422–433 (1997).
10. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
11. Lunetta, K. L., Boehnke, M., Lange, K. & Cox, D. R. Selected locus and multiple panel models for radiation hybrid mapping. *Am. J. Hum. Genet.* **59,** 717–725 (1996).
12. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286,** 964–967 (1999).
13. Mumm, S., Molini, B., Terrell, J., Srivastava, A. & Schlessinger, D. Evolutionary features of the 4-Mb Xq21.3 XY homology region revealed by a map at 60-kb resolution. *Genome Res.* **7,** 307–314 (1997).
14. Schwartz, A. *et al.* Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum. Mol. Genet.* **7,** 1–11 (1998).
15. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409,** 934–941 (2001).
16. Jobling, M. A. *et al.* A selective difference between human Y-chromosomal DNA haplotypes. *Curr. Biol.* **8,** 1391–1394 (1998).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank C. Nusbaum and T. Hudson for materials and advice on radiation hybrid mapping; J. Crockett, J. Fedele, N. Florence, H. Grover, C. McCabe, N. Mudd, S. Sasso, D. Scheer, R. Seim and P. Shelby for technical contributions; and D. Berry, J. Bradley, Y. Lim, A. Lin, D. Menke, M. Royce-Tolland, J. Saionz and J. Wang for comments on the manuscript. Supported in part by NIH.

Correspondence and requests for materials should be addressed to D.C.P. (e-mail: dcpage@wi.mit.edu).

Chapter 6

The Future of Genomic Mapping

Historians are fond of assigning "ages" to periods, which reflect the major technological forces that shaped human development during that time. The last few decades appear to be most popularly referred to as the "Space Age" in honor of the amazing progress mankind has made in escaping from our gravity well. While advancements in this arena have continued, it seems that exponential development has ended, and, even granted the growing presence of the International Space Station, further development of space-based resources will likely proceed very slowly, and without major impact on human society for the next several decades.

The Space Age has seen the concurrent growth of the computer industry. For the time computer technology, as measured by processor speed, still seems to be in exponential growth. It is probable that this growth will soon reach limits set at the quantum level. The Information Age will continue to flourish, however, in the form of massively parallel machines, or large networks of distributed processors. While these revolutions are happening currently, they are also becoming prosaic - the incredible processing power of modern systems are for many people solutions without a problem.

It is likely that the start of this century will be associated primarily neither with the Space or Information Ages, although the influences of the latter will likely continue to be important. Making predictions about the biological sciences is dangerous, given the failure of many such predictions to materialize. It seems clear, however, that molecular biology is at an unprecedented point in its development. Multiple forces have grown in weight to provide a critical mass that will raise molecular biology as one of the

hallmarks, perhaps *the* hallmark, of the current Age. The knowledge base in the field has grown to the point where it has become possible to describe and understand many fundamental cellular, and some medical, processes at the molecular level. The technology is available to accurately study biological compounds at concentrations approaching, or reaching, a single molecule. Widespread use of these techniques is dramatically reducing the material and economic costs of such studies. A trend towards miniaturization is allowing experiments once requiring dedicated facilities to be moved to the benchtop, and thus into small labs and even doctors' offices. Public enthusiasm for the field, or more appropriately for the cures and products that it has promised, has fueled an increase in both research funding and the ability to train more scientists with the appropriate skills. Finally, and most importantly, biological research is proving to be profitable, and looks likely to continue to be so for the foreseeable future.

It seems a little too pat to refer to the era as the Genomics Age, but there is a reasonable probability that the term will stick. Aside from being easily pronounced and with few syllables, "genomics" has become a household word recognized, at least in general form, by many non-scientists. Genomics will also clearly continue to play a role in the field, although with several complete genomes finished or near completion, more resources will likely now be shifted towards studying the biology downstream of DNA and RNA. But where does mapping fit into this new Age?

I will argue, perhaps ironically, that the era of genomic mapping, and possibly mapping in general, is at an end. Maps of sequence-tagged sites are totally irrelevant once the full sequence of an organism is in hand - the sequence contains all the information contained within the map, in addition to the nucleotide-level detail that a map will never provide. The Human Genome

Project (HGP) required maps to achieve its goal, and in this capacity genomic maps, including the map of the Y chromosome, have served wonderfully. The process selected by the HGP - generating markers, ordering them by mapping, then sequencing clones chosen from those maps - will ultimately produce the comprehensive sequence of the human genome at unprecedented precision. However, changes in technological paradigms have reduced the importance of the map, to the point that it is unlikely that maps will be produced for future genome projects.

A new genome sequencing effort currently has two general options. A whole genome shotgun, as pioneered by Celera, is clearly the method of choice for simple organisms with small, repeat-poor genomes. For more complex genomes, it is an attractive choice if completeness is not required, and if an understanding of the precise genomic organization is unnecessary. Whole genome shotguns are relatively cheap and rapid, as the development of markers, maps and clones are avoided. Regardless of the "forward-looking statements" made to Congress or in corporate quarterly reports, however, they suffer from an inability to fully and completely assemble complex, repeat-rich genomes.

This distinction is not related to maps, however - the deciding factor is the use of clones. Representing only a tiny fraction of higher genomes, BAC clones may be easily and rapidly assembled from shotgun sequence. The assembled BAC sequences may then themselves be assembled into the final, complete genomic sequence. The existence of a map, and corresponding STSs, certainly makes selection and ordering of BACs much easier. Generation of the markers and map represent a significant expenditure of resources, however, and are not necessary. The following scheme, proposed initially by the Hood lab as the "sequence-tagged-connector" scheme (Roach

et al., 1995; Siegel *et al.*, 1999), serves as an alternative for high precision genome sequencing and assembly:

A dense BAC library, on the order of 10-20x coverage of the genome, is constructed. Individual BACs are isolated and stored (for a 3Gb genome, with 250kb average insert size, this represents at most 250,000 BACs, or 650 384-well plates). BAC end sequences are obtained for each BAC, from both ends of the insert. A random selection of BACs (perhaps 20) is then shotgun sequenced. Comparison of each BAC's sequence to the BAC end library will then identify overlapping BACs. These are selected and sequenced to extend the growing contigs, then the process is repeated. Each contig will thus grow to form the sequence of a complete chromosome arm (or fuse to another contig in the process). As contigs finish, the pipeline can be kept in operation by choosing new BACs with novel BAC ends. In this fashion, an entire genome can be sequenced in the total absence of markers or maps.

Experience on the Y chromosome has indicated that highly-conserved, locally structured repeat families can represent snares for the unwary sequencer. Recognizing that these regions exist allows for their detection, however. Both palindromes and direct repeats contain three unique signature regions - the two termini of the repeat cluster, and the boundary between repeats. Identification of such sequences alerts the sequencer that a structured repeat has been entered, and allows the characterization of specific repeats through analysis of sequence family variants or other techniques (Kuroda-Kawaguchi *et al.*, 2001).

Clearly mapping will continue for some time yet, as many of the available resources have already passed the most laborious stages of development. Currently it is also possible that the human and material resources needed to direct a mapping effort will be calculated to be less

expensive than a shotgun or directed sequencing approach, especially for researchers working with limited funding. It is very unlikely, however, that full genomic mapping projects will be started from scratch on new organisms.

Maps have served their purpose in launching the era of genomics, and have performed wonderfully in their capacity. Without genomic maps, the human genome project would never have begun, and it is unlikely that commercial interests would have pondered a shotgun assembly of the genome without the publicly available maps and sequence as a scaffold. The passage of maps into obscurity is an exciting indication of the speed with which biological science and technology is progressing, and is not unique to this methodology (for example, restriction-length polymorphisms have effectively been totally replaced by simple sequence repeats and single-nucleotide polymorphisms in genetic studies). The challenge now is to use the experience and insights gained through mapping to develop the next generation of high-throughput genomic tools.

Reference

Kuroda-Kawaguchi T., Skaletsky H., Minx P., Brown L., Rozen S., Wilson R., Waterston R., and Page D. (2001). The DNA sequence of the human Y chromosome. *Unpublished.*

Roach J. C., Boysen C., Wang K., and Hood L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26:** 345-53.

Siegel A. F., Trask B., Roach J. C., Mahairas G. G., Hood L., and van den Engh G. (1999). Analysis of sequence-tagged-connector strategies for DNA sequencing. *Genome Res* **9:** 297-307.

Appendix A

A CD-ROM containing the Perl and C code used in the production of this thesis is enclosed with this manuscript. The files were edited with BBEdit on the Macintosh, and assume 4 characters per tab. The following files are included:

Software

`SimulateRH.cgi (Perl 5)`: The basic code for RH Simulator. Generates the front end, the graphics for each output, and calculates probability curves for models. It also requires the following packages:

`Probability (C/XS)`: A package containing the subroutines for performing likelihood calculations. The code itself is in `Probability.xs`, while the executable may be found buried in `blib/arch/auto/Probability/Probability.so`. You will need to modify the @INC designation in `SimulateRH.cgi` to point to your own local copy of the .so file (or more conveniently, to a symbolic link of that file).

`CalcTheta (C/XS)`: Fast subroutine for calculating theta scores. The code is derived from Perl code originally written by Helen Skaletsky. As for `Probability`, you will need to make modifications to @INC.

`ParseFasta.pl (Perl 5)`: Module for parsing ASCII FastA files.

`ParseQuery.pl (Perl 5)`: Module for extracting variable information from the location line of a browser (i.e., the information that is passed to a CGI script after the "?" in the URL).

`Other software/ (folder)`: Includes some code not discussed in this thesis, but might be of interest to a dedicated reader all the same. In particular `Fasta.cgi` is a very useful front end to the FastA algorithm, that includes some novel sequence complexity analysis algorithms.

Support Files

`order (ASCII)`: A simple list of the presumed order of markers along the Y chromosome.

`ModelTypes(ASCII)`: A collection of models. Models are designated in the file with linkage indicated with an asterisk (`*`), and non-linkage with a pipe (`|`).

`ModelStats(ASCII)`: Probability curve data generated for the above models. This data can be generated *de-novo* from within the program, for user specified models and distances.

`AllMatrices(ASCII)`: A list of all the vectors generated on the TNG panel. Marker name is followed by a tab, and then the vector.

`Data/ (Folder)`: Folder that `SimulateRH.cgi` references for some auxiliary files (included within).

`Thesis.pdf`: This document, in PDF format. About 9Mb in size.