Recombination patterns reveal information about centromere location on linkage maps

MORTEN T. LIMBORG,†‡ GARRETT J. MCKINNEY,* LISA W. SEEB* and JAMES E. SEEB*
*School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Box 355020, Seattle, WA 98195, USA, †National Institute of Aquatic Resources, Technical University of Denmark, Vejløvej 39, 8600 Silkeborg, Denmark

Abstract

Linkage mapping is often used to identify genes associated with phenotypic traits and for aiding genome assemblies. Still, many emerging maps do not locate centromeres – an essential component of the genomic landscape. Here, we demonstrate that for genomes with strong chiasma interference, approximate centromere placement is possible by phasing the same data used to generate linkage maps. Assuming one obligate crossover per chromosome arm, information about centromere location can be revealed by tracking the accumulated recombination frequency along linkage groups, similar to half-tetrad analyses. We validate the method on a linkage map for sockeye salmon (Oncorhynchus nerka) with known centromeric regions. Further tests suggest that the method will work well in other salmonids and other eukaryotes. However, the method performed weakly when applied to a male linkage map (rainbow trout; O. mykiss) characterized by low and unevenly distributed recombination – a general feature of male meiosis in many species. Further, a high frequency of double crossovers along chromosome arms in barley reduced resolution for locating centromeric regions on most linkage groups. Despite these limitations, our method should work well for high-density maps in species with strong recombination interference and will enrich many existing and future mapping resources.

Keywords: centromeres, genomic architecture, genotyping by sequencing, linkage mapping, recombination

Introduction

High-throughput sequencing technology has greatly accelerated the construction of dense linkage maps in nonmodel species (e.g. Amores et al. 2011; Miller et al. 2012; Everett & Seeb 2014). The emerging maps allow unprecedented insights into the genomic architecture of adaptive divergence in the wild (e.g. Chutimantsakun et al. 2011; Gagnaire et al. 2013; Richards et al. 2013). These studies catalyse our understanding of the number, genomic location and colocation of loci affected by natural selection. Unfortunately, centromeres are rarely included on linkage maps because the additional mapping efforts such as mapping half-tetrads add significant logistical hurdles (cf., Thorgaard et al. 1983; Brieuc et al. 2014).

Centromeres represent a fundamental component of chromosomal structure and function (Henikoff et al. 2001), and information about centromere location is vital for properly understanding genomes. Studies describing genetic divergence have shown striking patterns in either centromeric or telomeric regions (Carneiro et al. 2009; Ellegren et al. 2012). Hence, knowledge about chromosome type (i.e. acrocentric vs metacentric) is of paramount importance for understanding the underlying architecture of adaptive traits. However, interpretations often suffer from the lack of known centromeres on reference maps and genomes; this has impoverished interpretations of results plotted along high-density maps (e.g. Wang et al. 2012; Gagnaire et al. 2013; Carlson et al. 2015) or genome assemblies (e.g. Ellegren et al. 2012; Tine et al. 2014; Xu et al. 2014).

Here, we demonstrate a straightforward method to identify centromeric regions on linkage maps by phasing the same recombination data used to construct the map. We validate the method by comparing phased centromere placement with more direct centromere placement using half-tetrad analysis in sockeye salmon, Oncorhynchus nerka. Finally, we provide test examples that highlight advantages and limitations of the method for mapping centromeres in different sexes and taxa.

© 2015 The Authors. Molecular Ecology Resources published by John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
Materials and methods

Mapping data

We used genotype data from the mapping families presented in Limborg et al. (2015) to illustrate how consideration of recombination patterns in mapping data can be used to infer both chromosome type and location of centromeric regions. Two family data sets were generated from a single female sockeye salmon. Family 1 consisted of 77 gynogenetic diploid offspring genotyped for 2562 loci that were used to identify centromeric regions through half-tetrad analysis (Thorgaard et al. 1983). Family 2 consisted of a gynogenetic haploid family (n = 94) genotyped for 3245 loci that were used to construct a linkage map. Loci were scored using RAD sequencing and 5’-nuclease genotyping (Limborg et al. 2015).

Identifying chromosome type and centromeres

Here, we phased the data available from the haploid Family 2 to demonstrate an alternative approach for placing centromeres. Markers that are polymorphic within a mapping parent contain information about the allelic phase within that parent. After the construction of a linkage map, one can plot, for each offspring, the underlying phased genotypes along each linkage group (Fig. 1). This framework allows the observation of individual recombination events along linkage groups. We can then estimate a proxy for the recombination frequency between the terminal markers and each subsequent marker along a linkage group. We obtain recombination frequencies (RFm) for intervals between each marker (m) and the two terminal loci on a linkage group. Thus, along the entire length of each linkage group, we plot two patterns of RFm, one proceeding from each end, with each terminal marker assigned as a reference starting point (RFm = 0).

Interpretation of RFm

Certain assumptions are needed for the phasing method to work properly. First, strong recombination interference along chromosome arms should occur. Also, crossovers are needed on each side of the centromere in metacentric chromosomes. These assumptions are generally met in the sockeye salmon linkage map (Limborg et al. 2015) and lead us to expect exactly one crossover per arm in most meiotic products. Hence, the method should apply to any eukaryote genome characterized by strong recombination interference along chromosome arms.

It is important to consider the recombinant type of meiotic products captured in the form of haploid offspring. In haploid offspring, all four meiotic products are represented; however, almost always only two of the four chromatids on each side of the centromere engage in recombination during meiosis (Roeder 1997). We therefore expect ~50% of haploid gametes to represent meiotic recombinants for each chromosome arm. When scoring values of RFm along a linkage group, recombination events are recorded whenever a phase shift compared to the terminal reference marker (m0) is observed (Fig. 1a,b). With strong interference, this should translate into values of RFm ranging from 0.0 at m0 to ~0.5 when a full chromosome arm has been covered because only half the offspring will be recombinants.

![Fig. 1 Conceptual illustration of using phased genotypes to estimate recombination frequency (RFm). (a) Illustration of how to calculate RFm along a single chromosome arm from phased genotype data. The ten horizontal bars represent phased genotypes along a single chromosome arm for ten haploid individuals. Alternate phases are shown in red and blue. Starting at one terminal marker (m0) and moving along the linkage group, RFm record the fraction of offspring with an observed crossover (i.e. phase shift) between m0 and all subsequent markers (m1–6). (b) Resulting plot of the RFm values estimated in (a) above. When capturing all four meiotic products individually, as done in haploid mapping, RFm values are expected to vary from 0.0 at m0 to ~0.5 when a full chromosome arm has been covered because only half the offspring will be recombinants.](https://example.com/fig1.jpg)
the phase of the terminal reference marker is expected to be independent from phases of most markers on the arm opposite the centromere due to a general lack of cross-over interference along entire metacentric chromosomes (Colombo & Jones 1997; Demarest et al. 2011; Limborg et al. 2015). Therefore, for acrocentric chromosomes, RF\(_m\) will be 0.0 at the terminal reference marker and increases linearly towards a value of 0.5 at the opposite end (see pattern in Fig. 1b) – regardless of which reference marker represents the centromere. For metacentric chromosomes, plots of RF\(_m\) should translate into a flat pattern where RF\(_m\) plateau along the chromosome arm opposite the terminal reference marker reflecting random phasing on the opposite arm in relation to the reference marker (i.e. RF\(_m\) \sim 0.5).

By scoring RF\(_m\) from both ends of a linkage group, centromeres on metacentric chromosomes are then expected to reside in regions on the linkage group where values of RF\(_m\) counted from either end intersect around a value of 0.5. Plotting RF\(_m\) alone is thus expected to efficiently distinguish metacentric from single-armed (acro- or telocentric) chromosomes although RF\(_m\) alone will not identify the centromeric end on single-armed chromosomes. The resolution of most linkage maps in nonmodel species may prevent distinction between telocentric and acrocentric chromosomes (Levan et al. 1964). In the following, we assume that chromosomes always have a short p arm and refer to single-armed chromosomes as acrocentric (Levan et al. 1964). In conclusion, we expect plots of RF\(_m\) to broadly distinguish acrocentric from metacentric regions on metacentric chromosomes.

The paired families of Limborg et al. (2015) offered a perfect opportunity to compare the phasing approach of centromere mapping to known centromeric regions defined from the more direct inference from half-tetrad analyses.

**Application of RF\(_m\) to other example data sets**

We applied the phasing method to genotype data from published linkage maps that lacked information about centromere locations. First, we demonstrated the method in a map for pink salmon, O. gorbuscha (Limborg et al. 2014), based on gynogenetic haploid mapping data similar to that in haploid Family 2 for sockeye salmon presented above. We then applied the method to a male map for rainbow trout, O. mykiss (Miller et al. 2012), built from androgenetic doubled haploids. Lastly, we plotted RF\(_m\) patterns for a plant genome using the doubled haploid barley, Hordeum vulgare, family presented in Chutimanitsakun et al. (2011).

**Results**

We plotted paired RF\(_m\) estimates for each linkage group. Comparing plots of RF\(_m\) to centromeric regions previously defined in Limborg et al. (2015) revealed patterns that reflected acrocentric chromosomes (paired straight lines; Fig. 2a) or metacentric chromosomes (mirrored hockey stick shapes; Fig. 2b). This pattern was consistent across all linkage groups (Fig. S1, Supporting information). These results clearly illustrate the usefulness of considering RF\(_m\) to distinguish acrocentric from metacentric chromosomes in a female salmonid linkage map. For metacentric chromosomes, regions where RF\(_m\) estimated from either telomeric end converge around a value of 0.5 define centromeric regions with comparable power to the more direct inference from half-tetrad analyses (Figs 2 and S1, Supporting information).

Although our method worked well with data from a single female sockeye salmon, this does not mean that it will work well in all circumstances. Tests on linkage maps from other species with a more diverse set of
chromosomal characteristics revealed that our method works for some mapping data, whereas other data sets presented some limitations.

First, when we applied the method to the recently published map for pink salmon (Limborg et al. 2014), RFm patterns clearly depict a karyotype dominated by metacentric chromosomes (Fig. S2, Supporting information), consistent with the known pink salmon karyotype (Phillips & Rab 2001). This high-resolution map was based on a gynogenetic haploid family in a species with strong interference.

Second, the rainbow trout male map was characterized by generally short linkage groups and with many loci binned to the same map location. Consequently, we also observed a reduced resolution for the RFm method (Figs 3a and S3, Supporting information). At best, weak inference about chromosome type can be made for some of the larger, presumably metacentric, linkage groups (e.g. LGs WS01-WS08), but the RFm patterns remain inconclusive for most linkage groups in this male map.

Finally, in the barley map, most linkage groups were longer than 150 cM in accordance with reduced levels of interference; this resulted in patterns of wide regions characterized by RFm values of 0.5 because estimates from both sides plateaued after approximately 50 cM from either end (Figs 3b and S4, Supporting information). One linkage group (4H) was shorter (~130 cM), with putatively fewer double crossovers on each arm, and revealed a more satisfactory resolution for defining the putative centromeric region (Fig. S4, Supporting information). Estimates of RFm for all data sets are given in Table S1 (Supporting information).

Discussion

Here, we demonstrate a method to identify centromeres on metacentric chromosomes with similar accuracy to that obtained from half-tetrad analyses. Our RFm method provides a straightforward approach to improve many linkage maps by providing information on chromosome type and centromere location using existing data. Comparing the RFm method with half-tetrad data for sockeye salmon revealed high precision for locating centromeres on metacentric chromosomes. The method also provided convincing results for a previously published linkage map for pink salmon that lacked information about the location of centromeric regions. Lastly, although we validated the method for a haploid family, the method will also work in diploid families where both parents contribute alleles to offspring. In diploid families, estimates of RFm can be obtained in a similar fashion if restricting analyses to loci that only segregate in a single parent (e.g. ♀ = AB, ♂ = AA), where the allelic phase of the segregating parent can be unequivocally phased for estimating RFm.

Some analyses may require a concrete definition of centromeric regions (cf., Limborg et al. 2015), which will necessitate designation of a general threshold value of RFm to define centromeric locations. In our sockeye salmon example, defining a region flanked by the first markers reaching a value of RFm ~ 0.45 counted from each end seems to compare well with the regions defined based on half-tetrad analyses (Fig. S1, Supporting information). However, the optimal threshold value is likely to vary depending on the resolution and density of a given map.

Comparison to alternative methods

The phasing method offers a cost-effective and straightforward approach for including information about centromere locations on both existing and future linkage maps. Performing half-tetrad analyses to directly estimate gene–centromere distances is nontrivial. Historically, generation of mapping and half-tetrad families have been split into independent studies (e.g. Lindner

Fig. 3 Examples with data where the RFm method has limitations for locating centromeres. Graphs show RFm plotted along a linkage group for a male rainbow trout (a) and along a barley linkage group (b). RFm values are counted from the left (cM = 0; blue circles) and from the right (red circles). Both the rainbow trout and the barley linkage maps were constructed from doubled haploid families; RFm values are therefore expected to plateau around 0.5 (dotted horizontal line) when covering a full chromosome arm or when at least one crossover has occurred in each offspring.
et al. 2000; Morishima et al. 2001; Reid et al. 2007; Li et al. 2008) or, in a few cases, relied on large and expensive data sets including both types of crosses (Brieuc et al. 2014).

Other strategies to identify centromeres include cytogenetic methods targeting heterochromatin proteins known to encapsulate centromeres (Fransz et al. 2000; Kelly et al. 2006). However, this method also provides additional hurdles in terms of the expertise and analyses needed. The added complexity and cost of most alternative approaches means that many maps for nonmodel species may continue to lack information about centromere locations.

The RFm method can reduce the effort for placing centromeres on acrocentric chromosomes. Large families of half-tetrads were used in the studies of Brieuc et al. (2014) and Limborg et al. (2015) to obtain adequate resolution for locating centromeric regions. Considering that the resolution provided by the haploid mapping family to locate centromeres was similar to that obtained with the half-tetrads, we argue that a much smaller sample size of half-tetrads would suffice to simply distinguish centromeric ends from telomeric ends on acrocentric chromosomes. Alternatively, if centromere placement exists for a related species, or population, then centromeres on acrocentric chromosomes may be inferred by combining comparative mapping (cf., Kodama et al. 2014; McKinney et al. 2015) with information from RFm.

**Limitations of the method**

It should be noted that the RFm method may fail to detect centromeres on small metacentric chromosomes characterized by short arms (i.e. \(<\lesssim 50 \text{ cM}\)) where many meiotic products may completely lack observed crossovers along some arms. This is exemplified by the small metacentric linkage group S028 in the sockeye salmon map (Fig. S1, Supporting information).

In many species, one sex often shows reduced recombination rates (Lagercrantz & Lydiate 1995; Broman et al. 1998), which can lead to low resolution of linkage maps. Maps with low resolution are often characterized by having many loci binned to the same location, and this obstructs the resolution of RFm. In salmonids, for example, males exhibit lower recombination rates compared to females (e.g. Lien et al. 2011; Everett et al. 2012), and when we plotted RFm along a male linkage map for rainbow trout, signals of centromere locations were vague. Indeed, to further test the generality of such sex difference, we further estimated RFm for a diploid family of Chinook salmon (O. tshawytscha) using data from McKinney et al. (2015). Female recombination data provided an adequate resolution compared to the haploid female map for pink salmon, whereas the male map proved inadequate for inferring centromeric locations (data not shown). We conclude that for diploid crosses, RFm only appears reliable when the female parent, or the heterogametic sex, is used.

Furthermore, the RFm method will also have reduced resolution for chromosomes characterized by long arms (\(\gg 50 \text{ cM}\)) with incomplete interference such as in the barley example. In those cases, the value of RFm will level off after \(~50 \text{ cM}\) because the rate of double crossovers steadily increases when moving along an arm, obscuring signalling of the actual centromere location.

Hence, the method appears most promising for chromosomes with an average of one crossover per arm as seen in species with strong interference. Crossover interference is a common phenomenon, but varies between sexes (Drouaud et al. 2007), among species (Segura et al. 2013) and even among chromosomes within a genome (Lian et al. 2008). The method described here may work well on an array of taxa characterized by strong interference and short chromosomes such as birds (Tiersch & Wachtel 1991; Zhang et al. 2014), many mammals (Segura et al. 2013) and at least within the salmonid linkage among teleosts (Thorgaard et al. 1983; Lindner et al. 2000; Limborg et al. 2015); however, results may vary among species and testing is warranted.

**Future perspectives**

Centromeres are an essential component of the genomic landscape, and genetic maps lacking centromeres are incomplete. In humans, for example, subtelomeric regions are known to harbour evolutionary dynamic regions (Mefford & Trask 2002). Inference of similar patterns in other taxa requires information about the location of centromeres. Indeed, recent studies on speciation divergence report nonrandom distributions of divergence peaks in relation to centromeres and telomeres, patterns that suggest different roles of these genomic regions during adaptive divergence (Ellegren et al. 2012). Unfortunately, comparisons of centromeric with telomeric regions have often been impeded by missing, or equivocal, identification of centromeres on linkage maps (e.g. Gagnaire et al. 2013; Santure et al. 2013; Limborg et al. 2014) or on genome assemblies (e.g. Ellegren et al. 2012; Jones et al. 2012). Further, linkage maps are also crucial in guiding de novo genome assemblies (Lewin et al. 2009; Ellegren 2014; Tine et al. 2014), and linkage maps with centromeres will have an increased value in such efforts. The RFm method presented here can help mitigate existing limitations and improve linkage maps by providing a cheap and fast approach to distinguish acrocentric from metacentric chromosomes and locate centromeres on metacentric chromosomes.
Acknowledgements

We thank Carita E. Pascal for assistance with molecular analyses. Ryan K. Waples and Wesley A. Larson are thanked for insightful discussions during the work. We further thank Darin Combs and colleagues for providing assistance and hatchery facilities to conduct the family crosses at Issaquah Hatchery (WA, USA). MTL was supported by the Danish Council for Independent Research’s career programme Sapere Aude (Grant # 12-126687).

References


Carneiro M, Ferrand N, Nachman MW (2009) Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (Oryctolagus cuniculus). Genetics, 181, 593–606.


Kodama M, Brieuc MS, Devlin RH, Hard JJ, Naish KA (2014) Comparative mapping between Coho Salmon (Oncorhynchus kisutch) and three other salmons suggests a role for chromosomal rearrangements in the retention of duplicated regions following a whole genome duplication event. G3: Genes, Genomes, Genetics, 4, 1717–1730.


Lian J, Yin Y, Oliver-Bonet M et al. (2008) Variation in crossover interference levels on individual chromosomes from human males. Human Molecular Genetics, 17, 2583–2594.


M.T.L., L.W.S. and J.E.S. conceived the study. M.T.L. developed the method with input from G.M.K. M.T.L. drafted the manuscript, and all authors commented on earlier versions.

Data accessibility
The genotype data for the gynogenetic haploid and diploid sockeye salmon families are available from Limborg et al. (2015) at Dryad with doi: http://dx.doi.org/10.5061/dryad.q675s. All other data sets are available through the original publications: pink salmon (http://dx.doi.org/10.5061/dryad.pp43m; Limborg et al. 2014), rainbow trout (appendix S2 in Miller et al. 2012), barley (http://wheat.pw.usda.gov/ggpages/maps/OWB/; Chutimanitsakun et al. 2011), Chinook salmon (http://dx.doi.org/10.5061/dryad.j245; McKinney et al. 2015).

Supporting Information
Additional Supporting Information may be found in the online version of this article:

Table S1 Estimated RFm values for all loci on each linkage map for the different species analyzed in this study.