Gene Expression to Genetical Genomics

ORIGINAL RESEARCH

A Dynamic Neighboring Extension Search Algorithm for Genome Coordinate Conversion in the Presence of Short Sequence Duplications

Bin Zhu, Lu Jiang, George E. Liu

USDA, ARS, ANRI, Bovine Functional Genomics Laboratory, Beltsville, Maryland 20705, USA.
Email: george.liu@ars.usda.gov

Abstract: Oligonucleotide arrays are increasingly used in comparative genomic hybridization (CGH) to detect genomic copy number variation (CNV). The design of these arrays usually prefers uniquely mapped probes but routinely includes multiply mapped probes within a genome to maintain the high coverage and resolution. These duplicated probes could cause several limitations: besides their effects on the CNV calling, this kind of design also leads to the difficulty of converting genome coordinates between different genome assemblies. In this study, we tested over 385,000 probes for the genome coordinate conversion between two cattle genome assemblies and found out 33,910 (8.8%) of these probes cannot be uniquely mapped due to short sequence duplications. We also studied the distribution pattern of these short sequence duplications and discussed their potential impacts. Finally, we proposed and tested a dynamic neighboring extension search (DNES) algorithm to solve this conversion problem in order to facilitate a direct migration and comparison of array CGH results across different genome assemblies.

Keywords: bovine genome, short sequence duplication, MegaBLAST

Gene Expression to Genetical Genomics 2009:2 29–36

This article is available from http://www.la-press.com.

Introduction

Recent findings revealed that genomic structural variation, including copy number variation (CNV, i.e. large-scale insertions and deletions) is common in normal primates, rodents and dogs.[1–8] In humans, more than 6,000 CNV regions have been identified in normal individuals, and at least several hundreds are common in the population (as of 07/2009, http://projects.tcag.ca/variation/). Undoubtedly one of the major technological breakthroughs in the CNV discovery was the development of oligonucleotide microarrays for comparative genomic hybridization (CGH) experiments.[9,10] The oligonucleotide arrays are based on synthetic probes 45–85 bp in length, representing a tiling path along the entire genome.[11–13] It is generally accepted that the first and important step of array CGH technology is to design these representative probes. The uniqueness and uniformity of the probes are two important factors that will determine the accuracy and completeness of the subsequent CNV calling and analysis. Such design process was often conducted by computer tools using a genome assembly. These design algorithms usually prefer uniquely mapped probes but routinely include multiply mapped probes within a genome to maintain the high coverage and resolution.

With the continuing updates of the genome assemblies, DNA sequences and their genome coordinates can be rearranged to create a more accurate genome representation. The above-mentioned array design which allows duplicated probes could lead to the difficulty of migrating genome coordinates across assemblies, thus prevent the direct comparison of array CGH results based on different assemblies. Tools like LiftOver have been developed to facilitate genomes coordinate conversion across assemblies and species. Using the blat algorithm and premade coordinate conversion files like *over.chain*,[14] one can migrate coordinates from one assembly to another in a batch mode. However, LiftOver was not designed efficiently to handle these duplicated probes: even under optimized settings, the mapping results could be incomplete thus misleading. Instead of mapping short DNA sequence and converting genome coordinates, other related tools like MUMer[15,16] were designed for large-scale genome alignment and comparison and are generally computationally intensive and demanding.

In this study, we illustrated that the probe mapping uncertainty is due to short sequence duplication. We tested the genome coordinate conversion using MegaBLAST for over 385,000 probes between two cattle genome assemblies (Btau_3.1 and Btau_4.0) and found out about 33,910 (8.8%) of these probes cannot be uniquely mapped. In order to map these duplicated probes, we proposed and tested a dynamic neighboring extension search (DNES) algorithm to progressively use the flanking sequence information. We discussed the potential impacts of these short sequence duplications in the context of the CGH array design and the CNV calling and analysis.

Materials and Methods

Array CGH and probe design

As described previously,[11] the whole-genome array was designed incorporating ~390,000 oligonucleotides of variable length (45–75 bp) to achieve a melting temperature of 76 °C and a median probe spacing of 5.8 kb (http://www.nimblegen.com/products/cgh/other.html). Probes for the whole-genome arrays were selected based on their uniqueness (allowing up to 2 perfect hits) and base pair composition. Briefly, starting with DNA sequence masked for repetitive elements in the genome,[17] clusters of 10 probes were selected at 5,000-bp intervals throughout the cattle genome. The interval spacing within each cluster was 25 bp. Instead of a single fixed length, oligonucleotides were of variable length to achieve a target melting temperature of 76 °C. Probe lengths were constrained to be a minimum of 45 bp and a maximum of 85 bp. Optimal probes were selected by evaluating the 10 probes for each cluster using a composite quality score. This quality score is a sum of quality components (C_1 ... C_n) multiplied by their importance weights (W_1 ... W_n). The components for the whole-genome array were the two uniqueness measures: 1) a boolean measure of the base pair composition that fails oligonucleotides having runs of homopolymers, and 2) the variance from the optimal Tm target. The best oligonucleotide from each cluster was selected and incorporated into an array design using ArrayScribe array design software (NimbleGen). The selection of oligonucleotide probes was independently made using either Btau_3.1 or Btau_4.0 version of the cattle genome.

Uniqueness of probes was ascertained by determining the number of 45 mer matches (the minimal oligonucleotide length) of each oligonucleotide in the genome, as well as the average frequency of the windowed 15mer oligonucleotides that comprised the longer oligonucleotide. However, they allow probes to have multiple perfect matches within the genome assembly when unique probes were not available at the desired density.

Difference between bovine genome Btau_3.1 and Btau_4.0

Btau_3.1 (August 2006) was produced with a combination of WGS and BAC sequence by the Atlas genome assembly system using a version of the Integrated Bovine Map that represents merged data from several independent maps.[18] Btau_4.0 is a newer (October 2007) assembly of the cattle reference genome (http://www.hgsc.bcm.tmc.edu/project-species-m-Bovine.hgsc?pageLocation=Bovine). This assembly added relatively little new sequence data, and thus contigs and scaffolds are not significantly changed, but used the ILTX and BAC finger-print contig maps and split scaffolds based on consistent bovine and sheep BES data to place the contigs and scaffolds in the genome, thereby resulting in more accurate chromosome structures.[19] Detailed description of genome assembly process was presented previously.[20]

Btau_3.1 to Btau_4.0 probe conversion

All the Btau_3.1 probes (385,048) were mapped using MegaBLAST against the Btau_3.1 and Btau_4.0 assemblies with minimum 40 base pair match required[21] (Parameters: -D 2 -v 7 -b 7 -e 1e-40 -s 90 -W 12 -F F). After initial search, only the matched results with 100% identity for the full length of the query sequence were kept. A total of 33,912 probes (8.8%) were found out having duplicated mapping locations in the Btau_4.0 assembly. For the comparison purposes, a sequence similarity search was also repeated on the Btau_3.1 assembly, and almost the same number of probes (33,910) was found to be duplicated mapped in this self-mapping exercises.

Results and Discussion

Mapping probes on genome assemblies

We first tested LiftOver to convert 385,048 probes designed using the Btau_3.1 assembly. Only 319,140 probes are successfully mapped onto the Batu_4.0 assembly under the LiftOver's default setting. A total of 65,908 probes (i.e. 17.2%) did not produce mapping results, which include both uniquely and multiply mapped probes. Using MegaBLAST,[21] we processed the same 385,048 Btau_3.1 probes in order to map them onto the Btau_4.0 assembly. In this search, only the matched results with 100% identity for the full length of the probes are kept. A total of 33,912 probes (~8.8%) were mapped to more than one location on the Btau_4.0 assembly. In order to test whether such duplicated hits are due to the difference between these two assemblies, we also used MegaBLAST to map these 385,048 probes against the Btau_3.1 assembly itself, and a similar result (33,910 probes, ~8.8%) were obtained. As shown in Table 1, both Btau_3.1 and Btau_4.0 assemblies have almost identical numbers of duplicated probes which strongly indicates that the duplicated probes are not due to the differences between the Btau_3.1 and Btau_4.0 assemblies. On the other hand, it suggested that the Btau_3.1 and Btau_4.0 assemblies produce almost identical results under our current MegaBLAST search setting. In fact, the discrepancy of 2 probes (the difference between 33,912 and 33,910) supported the description that the Btau_4.0 assembly adds relatively little new sequence data, and thus contigs and scaffolds are not significantly changed. The major difference between these 2 assemblies is that the Btau_4.0 assembly used different map information from that was used for the Btau_3.1 assembly to place the contigs and scaffolds in the genome, which resulted in more accurate chromosome structures.

However, these macro structural changes did not significantly interfere our mapping effort performed at the micro structural level (45–75 bp).

Dynamic neighboring extension search (DNES) algorithm

Since not all Btau_3.1 probes can be uniquely mapped to the Btau_4.0 assembly due to short sequence duplication, we proposed a dynamic neighboring extension search (DNES) algorithm to uniquely map these duplicated probes. The algorithm of our proposed method is presented in Figure 1. It includes several steps as follows:

Step 1: Use MegaBLAST to map the Btau_3.1 probes on the Btau_4.0 assembly.

Table 1. Duplicated probes mapped in the Btau_3.1 and Btau_4.0 assemblies.

	Btau_3.1	Btau_4.0
Number of duplicated probes	33,910	33,912
Total probes studied	385,048	385,048
Percentage of duplicated probes	8.8%	8.8%

Step 2: Count their mapping times on the Btau 4.0 assembly.
Step 3: If the mapping result from Step 2 is equal to 1, output mapped probe and go to next probe until all of them are processed. If it is greater than 1, retrieve the probe 's flanking sequences (50 bp at the both ends) and go back to Step 1 to repeat MegaBLAST with the elongated sequence. Iteratively increase the input sequence by retrieving flanking sequences for each cycle until the mapped result is equal to 1. Then output mapped probe and go to next probe until all of them are processed.

With this algorithm, all the duplicated probes were uniquely mapped on the Btau_3.1 and Batu_4.0 assemblies. Using this genome coordinate conversion, the CGH results based on the Btau_3.1 were successfully migrated onto the Btau_4.0 assembly.

Distribution landscapes of short sequence duplication

We further analyzed the distribution patterns for these duplicated probes. Depending on their mapped locations, the duplicated probes can be divided into

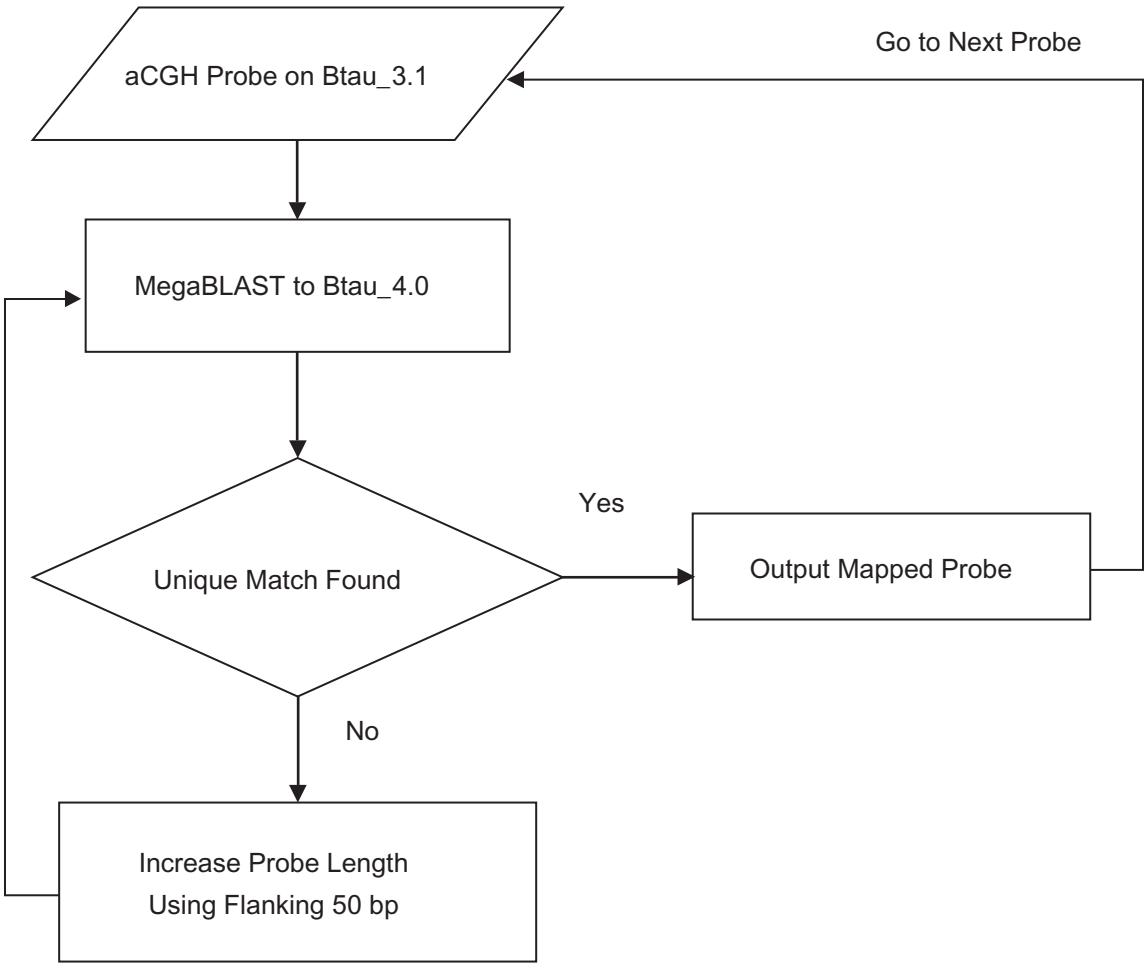

Figure 1. The flow chart of the dynamic neighboring extension search (DNES) algorithm. MegaBLAST is used to map the Btau 3.1 probes on the Btau 4.0 assembly. If the mapping result is 1, output mapped probe and go to next probe. If it is greater than 1, retrieve the probe's flanking sequences (50 bp at the both ends) and perform MegaBLAST with the elongated sequence. Iteratively increase the input sequence by retrieving flanking sequences for each cycle until the mapped result is equal to 1. Then output mapped probe and go to next probe until all of them are processed.

intrachromosmal and interchromosomal duplications. For intrachromosomal duplications, the majority of mapped locations exist on the same chromosome as tandem duplications, i.e. identical sequences in close proximity with either the same or inverted orientation. For interchromosomal duplications, mapped locations exist on different chromosomes. Some of these examples are illustrated in Figure 2. In panel A, two pairs of paralleled green lines represent direct tandem duplications. In panel B, one pair of paralleled green lines and two crossed red lines indicate inverted tandem duplications. In panel C, pairs of interchromosomal duplications are located at the different chromosomes. To further evaluate the pattern of short sequence duplication, we prepared a whole-genome distribution map of duplicated probes on the Btau_4.0 assembly (Fig. 3). To quantitatively evaluate the distribution of duplicated probes, the probe densities with respect to each chromosome are plotted in Figure 4. This probe density is computed as the number of duplicated probes in each chromosome divided by number of probes in the chromosome. A similar plot was obtained when the number of duplicated probes was normalized by the length of each chromosome (data not shown). The density distribution in both autosomal chromosomes and X chromosome is quite consistent (around 5%). The duplicated probes have larger density value in unassigned chromosome (chrUn) than the other chromosomes, which is expected since chrUn usually includes more uncertain or duplicated sequences than

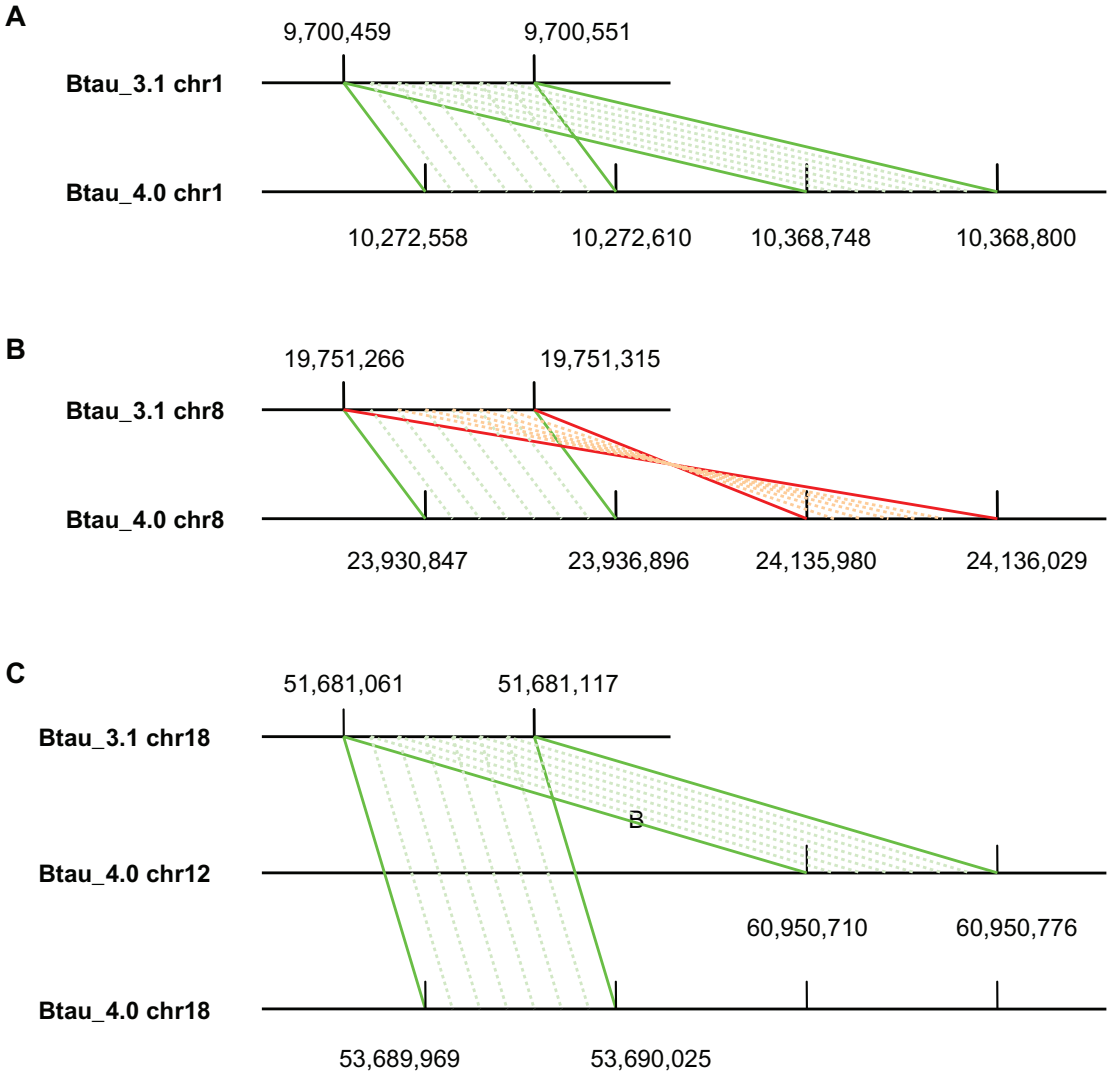

Figure 2. Examples of short sequence duplications. **A**) Two pairs of paralleled green lines represent direct tandem duplications; **B**) One pair of paralleled green lines and two crossed red lines indicate inverted tandem duplications; and **C**) Interchromosomal duplications are located at the different chromosomes.

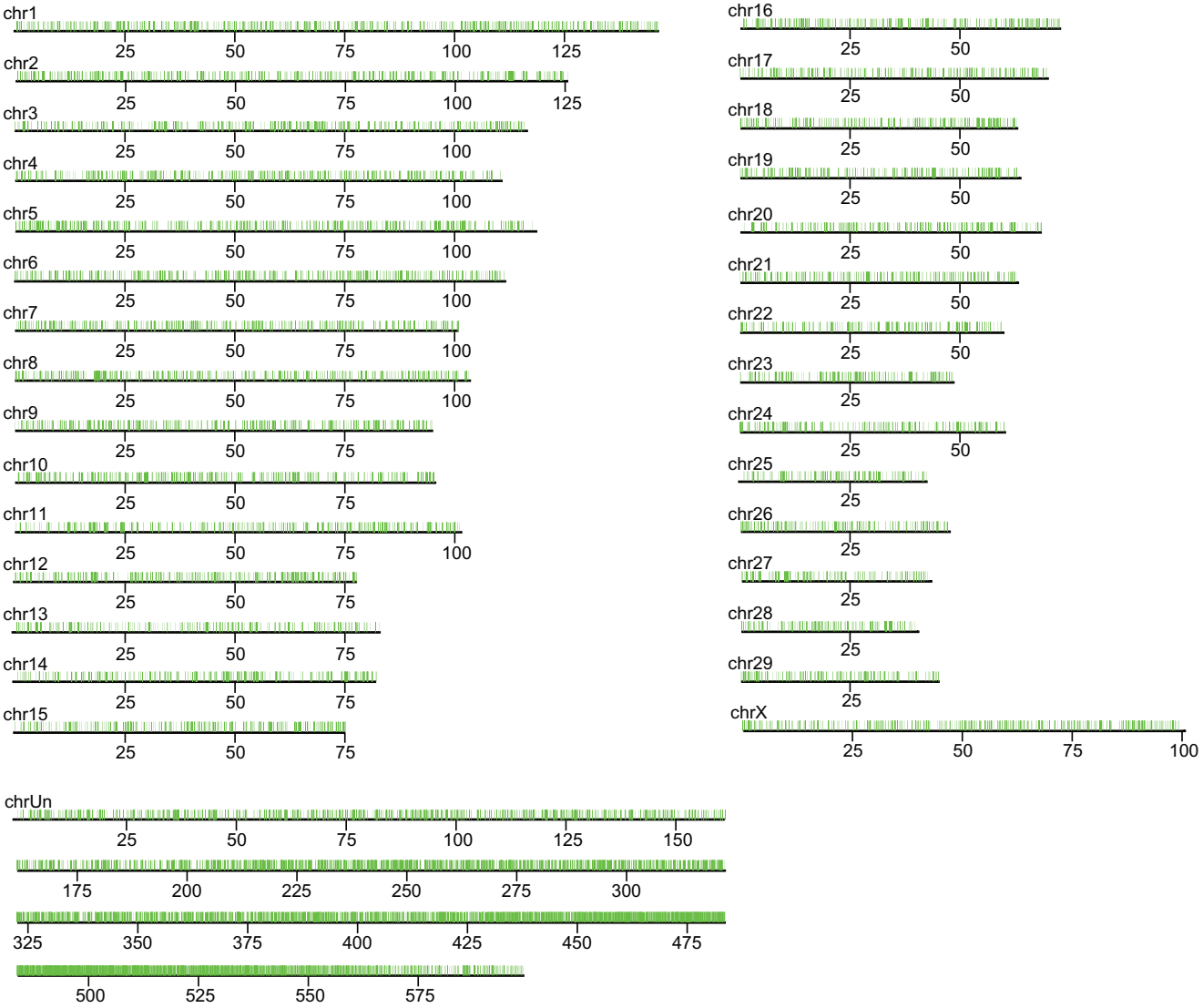

Figure 3. A genome-wide distribution map of short sequence duplications in the Btau_4.0 assembly. Short sequence duplications are represented as green bars on chromosomes and chromosomes are shown in the scale of Mb (1,000,000 bp).

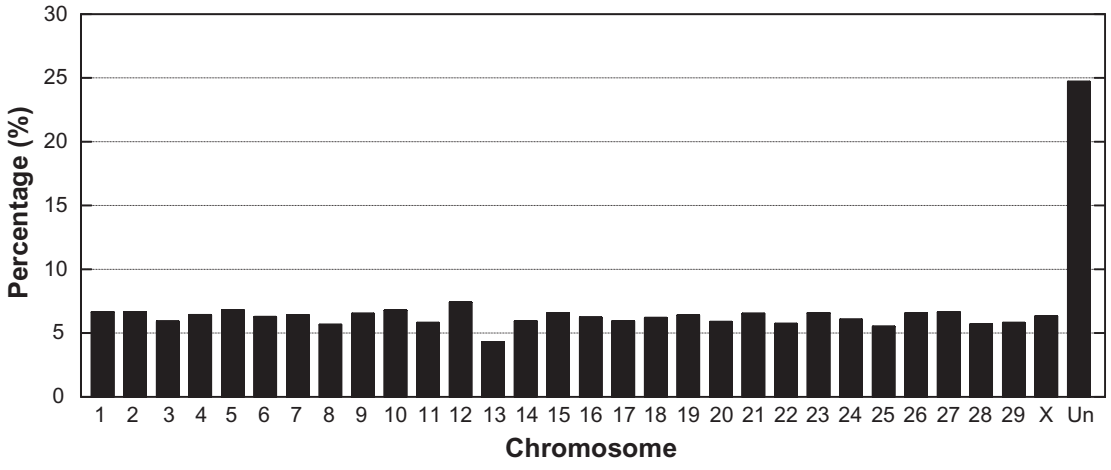

Figure 4. Duplicated probe density in each chromosome. The probe density is represented by the percentage of duplicated probes on each chromosome, i.e. the number of duplicated probes divided by the total number of probes in the chromosome.

the other chromosomes. All the duplicated probes seem to be uniformly distributed on the Btau_4.0 assembly, which suggested they are widely spread across the genome. The location variety shows complex nature of the short sequence duplication and therefore could propose challenges to the array CGH calling algorithm and CNV analysis if not taken into consideration. Besides the difficulties of genome coordinate conversion, these perfectly duplicated probes could cause several limitations: such as in CGH, hybridization signals on these duplicated probes are actually derived from multiple genomic loci which share the same or similar sequence. If not taken into consideration, these averaged values from distinct loci could confuse the CNV calling algorithms.

In this study, we began with a practical probe conversion task to facilitate the array CGH result migration and comparison across assemblies. We detected the widely spread short sequence duplication and proposed a dynamic neighboring extension search (DNES) algorithm to uniquely map duplicated probes. We studied their types and distribution patterns in the Btau_4.0 genome assembly. It is worthwhile to note that some of these duplications can be artificially created by the assembly error. An alternative assembly of the cattle genome (UMD2) was reported to be significantly improved and different from the Btau_4.0 assembly.[22] Comprehensive comparison of cattle genome assemblies prepared by different assemblers thus assembling algorithms is clearly beyond the scope of this study. We did map these probes on the UMD2 genome assemblies. We found that a fewer but significant portion (about 9,000 probes) has multiple mapping locations and around 8,500 probes cannot be mapped on the UMD2 assembly. Therefore, in conclusion, although their extent varies across different assemblies, these short sequence duplications are widely distributed. This short sequence duplication is significantly shorter than the well-studied segmental duplication, which by definition are referring to duplicated DNA segments with ≥1 kb length and ≥90% sequence identify.[19] In human, segmental duplications are often found to be organized around 'core' duplicons, which are enriched for genes and transcripts.[23] They function as hotspots for nonallelic homologous recombination leading to genomic disorders, copy number variation and gene and transcript innovations. Although, the short sequence duplication may overlap with segmental duplication occasionally, their generation and evolution are not yet understood and deserve our further investigation.

Disclosures

The authors report no conflicts of interest.

References

1. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004;305:525–8.
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature*. 2006;444:444–54.
3. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired- end mapping reveals extensive structural variation in the human genome. *Science*. 2007;318:420–6.
4. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*. 2008;453:56–64.
5. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, et al. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet.* 2007; 3:e3.
6. Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, Cook S, et al. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet*. 2008;40:538–45.
7. Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM. The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res*. 2009;19:491–9.
8. Chen WK, Swartz JD, Rush LJ, Alvarez CE. Mapping DNA structural variation in dogs. *Genome Res*. 2008.
9. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*. 1998;20:207–11.
10. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*. 1996;6:639–45.
11. Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, Nair P, et al. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer*. 2005;44:305–19.
12. Baumbusch LO, Aaroe J, Johansen FE, Hicks J, Sun H, Bruhn L, et al. Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics*. 2008;9:379.
13. Hester SD, Reid L, Nowak N, Jones WD, Parker JS, Knudtson K, et al. Comparison of comparative genomic hybridization technologies across microarray platforms. *J Biomol Tech*. 2009;20:135–51.
14. Kent WJ. BLAT-the BLAST-like alignment tool. *Genome Res*. 2002;12:656–64.
15. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. *Nucleic Acids Res*. 1999;27:2369–76.
16. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*. 2002;30:2478–83.
17. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics and Development*. 1999;9:657–63.
18. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*. 2004;428:493–521.
19. The Bovine Genome Sequencing and Analysis Consortium: The Genome Sequence of Taurine Cattle: a window to ruminant biology and evolution. *Science.* 2009;324:522–8.

20. Liu Y, Qin X, Song XZ, Jiang H, Shen Y, Durbin KJ, et al. Bos taurus genome assembly. *BMC Genomics*. 2009;10:180.
21. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol*. 2000;7:203–14.
22. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol*. 2009;10:R42.
23. Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet*. 2007;39:1361–8.