European Journal of Human Genetics (2019) 27:1436–1444
https://doi.org/10.1038/s41431-019-0419-0

ARTICLE

Heterogeneity in the extent of linkage disequilibrium among exonic, intronic, non-coding RNA and intergenic chromosome regions

Alejandra Vergara-Lope[1] · Sarah Ennis[1] · Igor Vorechovsky[1] · Reuben J. Pengelly[1] · Andrew Collins[1]

Received: 13 December 2018 / Revised: 4 March 2019 / Accepted: 16 April 2019 / Published online: 3 May 2019

Abstract

Whole-genome sequence data enable construction of high-resolution linkage disequilibrium (LD) maps revealing the LD structure of functional elements within genic and subgenic sequences. The Malecot–Morton model defines LD map distances in linkage disequilibrium units (LDUs), analogous to the centimorgan scale of linkage maps. For whole-genome sequence-derived LD maps, we introduce the ratio of corresponding map lengths kilobases/LDU to describe the extent of LD within genome components. The extent of LD is highly variable across the genome ranging from ~38 kb for intergenic sequences to ~858 kb for centromeric regions. LD is ~16% more extensive in genic, compared with intergenic sequences, reflecting relatively increased selection and/or reduced recombination in genes. The LD profile across 18,268 autosomal genes reveals reduced extent of LD, consistent with elevated recombination, in exonic regions near the 5′ end of genes but more extensive LD, compared with intronic sequences, across more centrally located exons. Genes classified as essential and genes linked to Mendelian phenotypes show more extensive LD compared with genes associated with complex traits, perhaps reflecting differences in selective pressure. Significant differences between exonic, intronic and intergenic components demonstrate that fine-scale LD structure provides important insights into genome function, which cannot be revealed by LD analysis of much lower resolution array-based genotyping and conventional linkage maps.

Introduction

The genome-wide pattern of linkage disequilibrium (LD) reflects the combined impacts of recombination, natural selection, genetic drift and mutation. Therefore, analysis of fine-scale LD structure provides opportunities to increase understanding of these important processes and their impacts on the genome. Previously, LD analysis has enabled the development of cost-effective genome-wide association studies, and the consequent mapping of numerous common disease genes, through development of arrays of 'tag' SNPs [1]. LD studies have also increased understanding of population structure and migration [2, 3], the nature of recombination hot-spots and the identification of sequence determinants which promote recombination [4, 5].

The ability to undertake cost-effective high sequence quality whole-genome sequencing (WGS), enables analysis of the properties of genomes at high resolution. Pengelly et al. [6] demonstrated that LD maps from WGSs yield ~2.8-fold as many regions of intense LD breakdown (which align with recombination hot-spots) compared with array-based tag genotypes, which miss substantial information. The increased resolution from sequence-based LD maps may provide further insights into the processes of selection and recombination operating at the gene and subgene levels. Furthermore, because the reliable recognition of disease-related variation in patient sequence data is challenging, increased understanding of the impact of recombination and selection at the genic and subgenic level [7, 8] may aid the prioritisation of candidate genes and variants.

LD maps based on the Malecot–Morton model [9–11] combine pairwise association data between single-nucleotide polymorphisms (SNPs) to quantify the variable rate of decline of LD with distance across SNP intervals. LD map distances are additive and analogous to the linkage

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-019-0419-0) contains supplementary material, which is available to authorised users.

✉ Andrew Collins
arc@soton.ac.uk

[1] Human Genetics, Faculty of Medicine, University of Southampton, Duthie Building (808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK

map centimorgan (cM) scale, but expressed in LD units (LDUs) where one LDU is the (highly variable) physical distance along the chromosome over which LD declines to 'background' levels. Plots of the LDU scale, compared with the physical kilobase (kb) maps show 'steps' where LD breaks-down over narrow sequence intervals (often aligning with recombination hot-spots) and 'plateaus' where LD is strong (regions, which align with 'blocks' of low haplotype diversity [1]). Earlier construction of genome-wide LD maps using this approach for array-based data [11, 12], including HapMap phase II [13], indicated that the genome of the CEU population (Utah Residents with Northern and Western European ancestry) has 57,819 LDUs. Given that the genome sequence spans ~3,100,000 kb the average extent of LD (kb/LDU) is ~54 kb. However, this figure is based on data from HapMap tag SNP arrays, which have much lower resolution than WGS [6]. The increased resolution from WGS-derived maps enables analysis of the LD structure of much finer-scale genomic features such as gene exonic and intronic sequences.

Previous analyses have considered the extent of recombination and LD within and between genes. McVean et al. [14] found recombination rates to be higher in intergenic regions close to genes, compared with recombination rates within genes. Eberle et al. [15] found a significant increase in the extent of LD in genic versus intergenic regions, which could not be explained purely by differences in the recombination rates, consistent with increased selection in genic regions.

Kong et al. [16] describe a fine-scale recombination map with 10-kb resolution. Using 10-kb bins classified as genic, intergenic, or at gene boundaries, they demonstrated reduced recombination rates in genic, compared with intergenic, regions along with some sex-specific differences. They noted lower recombination rates in bins containing only exons and higher rates for bins containing only introns, particularly for intronic bins distant from exons. Similarly, in intergenic regions, recombination rates were found to increase with distance from exons. For intergenic regions close to genes they observed reduced recombination closer to the 5′ ends of genes than to the 3′ ends. Bins containing the first exon of a gene were found to have a higher recombination rate than the last.

Berger et al. [17] found approximately 13.6% more LD in genic, compared with non-genic, regions of the genome in their study based on array-based genotyping (684,990 SNPs). However, their results do not correct for the substantial chromosome size-dependent differences in the average extent of LD, which reflect the higher recombination rates of smaller chromosomes [11].

We describe LD maps of the autosomal genome constructed from a large WGS data sample from individuals in the Wellderly study [18] (https://genomics.scripps.edu/browser/) of healthy, elderly, individuals aged >80 years from the general US population sequenced at high depth on the Complete Genomics platform. The much increased resolution of LD structure enables analysis of LD patterns on a very fine scale at the level of individual gene exons providing novel insights into the impact of recombination and selection on genome structure and function.

Materials and methods

Samples used and SNP processing

SNP genotypes were obtained from WGS data from the Scripps Wellderly Genome Resource comprising 454 unrelated individuals of ethnically European origin from the Wellderly study [18]. Following Pengelly et al. [6], we excluded SNPs with >5% missing genotypes and SNPs with a Hardy–Weinberg deviation p-value of <0.001 [19]. As rare SNPs are uninformative for LD, we evaluated the impact of excluding SNPs with alternative minor allele frequencies (MAF) of <0.05 and <0.01, using chromosome 22 as an example (Supplementary Fig. 1). We found both LD maps were very similar but using a <0.01 MAF cut-off produced a 3.4% longer map. Using a MAF <0.01 cut-off retains many more SNPs (103,367, compared with 70,579 SNPs retained using the MAF <0.05 cut-off) and may help better resolve LD structure in genomic regions with higher recombination rates. We therefore used all SNPs with a MAF of 0.01 or greater for subsequent work. The completed LDU maps of chromosomes 1–22 contain 7,162,973 SNPs (Table 1) spanning a total chromosome length of 2,791,110 kb indicating a density of one SNP every ~400 base pairs.

LD map construction

We undertook the construction of LD maps in LDUs for the autosomal chromosomes 1–22 using the LDMAP program, which implements the Malecot–Morton model. The program constructs maps iteratively using composite likelihood [9, 11, 13]. LDMAP evaluates the rate of decline of LD (parameterised as ε), in each interval between adjacent SNPs, using a sliding window, which weights association data from all SNP pairs in the region, which include the interval of interest. The corresponding LDU distance for the interval is εd where d is the physical distance in kilobases and LDU distances are additive to form a map contour (Supplementary Fig. 2).

LDU map analysis

We compared lengths of maps in LDUs with genetic map lengths in cM for chromosomes [20] (Table 1). We also

Table 1 Characteristics of whole chromosome maps

Chromosome	Chromosome start location (kb)	Chromosome end location (kb)	Chromosome kb coverage[a]	Chromosome LDU length	Number of SNPs	Chromosome length (cM)	LDU/cM
1	69.51	249,222.53	249,153.02	5078.92	557,873	270.27	18.79
2	11.94	239,856.97	239,845.03	4736.82	593,868	257.48	18.40
3	60.20	197,880.78	197,820.58	4138.09	509,066	218.17	18.97
4	13.26	191,033.02	191,019.76	3936.59	504,243	202.8	19.41
5	13.33	1,807,165.00	180,702.67	3785.07	459,987	205.69	18.40
6	148.00	170,919.74	170,771.73	3604.75	472,261	189.6	19.01
7	21.95	159,127.02	159,105.07	3460.39	415,335	179.34	19.30
8	161.47	146,296.84	146,135.37	3101.18	400,025	158.94	19.51
9	62.10	141,102.87	141,040.77	2953.02	311,320	157.73	18.72
10	92.19	135,506.38	135,414.19	3140.77	367,619	176.01	17.84
11	189.67	134,945.77	134,756.10	2943.56	351,378	152.45	19.31
12	83.15	133,838.99	133,755.84	2990.16	345,765	171.09	17.48
13	19,168.01	115,108.80	95,940.79	2309.50	261,818	128.6	17.96
14	19,050.28	107,288.38	88,238.10	2158.42	239,704	118.49	18.22
15	20,010.01	102,486.12	82,476.10	2151.48	207,177	128.76	16.71
16	83.89	90,180.71	90,096.83	2562.56	229,203	128.86	19.89
17	0.83	81,153.78	81,152.95	2287.03	195,607	135.04	16.94
18	11.28	78,015.56	78,004.28	2079.02	208,014	120.59	17.24
19	94.62	59,097.93	59,003.31	1869.27	163,978	109.73	17.04
20	61.10	62,964.27	62,903.17	1846.33	166,816	98.35	18.77
21	9495.96	48,100.71	38,604.75	1110.60	99,550	61.86	17.95
22	16,054.80	51,223.99	35,169.19	1184.17	102,366	65.86	17.98
Totals/ chromosome mean	-	-	2,791,109.60	63427.68	7,162,973	3435.71	18.36

[a]Table includes all heterochromatic and centromeric regions except acrocentric p-arms, which were not sequenced. Genome reference sequence hg19 was used throughout

compared LD structure with recombination rates as LDU/cM for chromosomes (Table 1).

We determined the extent of LD as kb/LDU for intergenic regions and also for non-coding RNA regions, genes and exons and introns within genes. The boundaries of gene, exon, intron, intergenic and non-coding RNA regions were identified using the NCBI RefSeq gene definitions. However, we recognise that the definition of genomic features is complex because, for example, there is variable exon utilisation in different transcripts and so discrimination between alternative genome features is far from straightforward. Custom scripts were used for linear interpolation to convert the sequence positions of the boundaries of these features into corresponding locations on the LDU map. Although the LDU maps are not linear, the use of linear interpolation for analysis of high-resolution maps is justified over short distances. We determined LDU locations and matched with approved gene names for 18,268 autosomal genes. For analysis of genic and intergenic regions, we followed Berger et al. [17] such that all genes, which overlap with other genes, were merged into a smaller number of 'genic regions'. A custom python script was implemented to establish the boundaries of genic regions. Intergenic regions were taken as any areas flanked by, but not overlapped by, genic regions. Heterochromatic regions from acrocentric chromosome p-arms were not included in the maps or subsequent analyses. All centromeric intervals (which also include centromeric heterochromatin) between the last gene on chromosome p-arms and the first gene on the q-arms of non-acrocentric chromosomes were excluded from analysis of intergenic regions because of the distinct properties of these regions. The LDU boundaries of all annotated exons and introns were determined by linear interpolation in the same way and genes partitioned into those transcribed on either the forward or reverse strand. The latter enabled unified analysis of all genes in the 5′ to 3′ direction. The small number of exons and introns involved in production of transcribed products on both forward and reverse strands were excluded from the exon/intron analysis. Non-coding RNA data were also clustered where overlapping but were not distinguished from other genomic features (such as our definition of intergenic regions) if they

Table 2 Extent of LD in kb (kb/LDU) in different genome regions

Chromosome	Whole chromosomes[a]	Genic regions	Gene exons	Gene introns	Non-coding RNAs	Intergenic regions[b]	Centromeric regions	Gene exons + non-coding RNAs
1	49.06	48.15	50.85	47.74	52.20	41.92	3772.89	51.96
2	50.63	59.96	63.60	60.02	57.00	44.34	262.41	57.79
3	47.80	49.39	50.65	49.15	52.80	45.19	413.50	52.51
4	48.52	51.26	53.97	51.43	48.65	46.11	1687.53	49.35
5	47.74	50.95	59.08	50.85	52.01	44.67	1331.94	52.78
6	47.37	50.05	47.72	49.96	52.53	44.01	445.44	52.83
7	45.98	48.38	45.15	48.42	54.39	41.63	383.35	52.83
8	47.12	48.42	54.20	48.48	53.04	44.20	377.97	53.16
9	47.76	43.88	42.31	43.79	48.19	38.89	1198.94	47.04
10	43.11	46.68	51.49	46.28	48.76	38.16	740.88	49.15
11	45.78	45.78	48.13	45.63	46.40	44.32	626.28	46.75
12	44.73	48.13	44.01	48.34	51.95	40.05	632.59	49.98
13	41.54	44.36	48.42	44.22	47.10	40.38	-	47.23
14	40.88	49.18	37.91	50.40	50.54	36.45	-	47.8
15	38.33	46.47	52.12	46.88	46.03	32.71	-	46.83
16	35.16	32.47	46.79	31.43	34.04	28.54	664.07	36.4
17	35.48	39.69	39.22	39.45	35.78	29.22	654.01	36.68
18	37.52	36.62	32.07	37.00	48.13	35.65	278.98	45.11
19	31.56	31.42	33.69	31.09	33.77	27.24	139.52	33.74
20	34.07	36.73	32.49	36.47	38.74	28.79	980.69	37.18
21	34.76	32.69	44.70	32.19	29.58	35.06	-	30.63
22	29.70	39.24	41.95	39.62	41.75	23.60	-	41.79
Chromosome means/SD	42.03/6.40	44.54/7.23	46.39/8.20	44.49/7.42	46.52/7.61	37.78/6.81	858.29	46.34/7.27

[a]Includes centromeric regions

[b]Excludes centromeric regions

overlapped these regions. The physical size of annotated features is given in Supplementary Table 1, the LDU size of corresponding features is given in Supplementary Table 2 and the corresponding counts of annotated features are given in Supplementary Table 3. To quantify the extent of LD for each feature, we used the ratio kb/LDU throughout which represents the extent of LD in kilobases for any genomic region (Table 2).

Variation in the extent of LD across genes

We examined the profile of variation in extent of LD across all genes, considering exonic and intronic regions separately. Because gene size is highly variable, we divided all genes into five bins oriented from 5′ to 3′, with bins equally sized for a given gene. The location of the mid-point in the sequence of each exon and/or intron was used to sum the LDU and kb length of that exon or intron into the respective bin (Supplementary Table 4). To examine the impact of highly variable gene size, we constructed LD extent profiles using the set of 18,268 genes divided into two groups of 9134 genes each corresponding to 'small genes' of size <23.5 kb and 'large genes' of size >23.5 kb (Supplementary Tables 5 and 6).

Variation in extent of LD for different gene groups

We examined the extent of LD for annotated genes by assigning them, where names and locations matched, to one of the five gene groups defined by Spataro et al. [21]. The classification is useful for examining the extent of LD and relationship to gene essentiality and disease. The gene groups are defined as:

1. Essential non-disease (END) genes, 1572 putatively essential genes defined as orthologues of mouse essential genes detected by knock-out experiments and not involved in any human disease. 2. Non-disease non-essential (NDNE) genes, 13,135 genes not known to be involved in any human disorder and not known to be essential. 3. Complex non-Mendelian (CNM), 2388 genes uniquely associated with complex

diseases. 4. Complex-Mendelian (CM), 203 genes associated with both complex and Mendelian disorders. 5. Mendelian non-complex (MNC), 684 genes uniquely causing Mendelian disease traits. We determined the extent of LD in each of the five gene groups (Supplementary Table 7).

Results

Whole chromosome LDU maps and comparison with linkage maps

The LD map of the autosomes (Table 1) has ~63,428 LDUs, which is of similar magnitude to earlier estimates from the CEU population using HapMap phase II data (but which also included the X chromosome) of 57,819 LDUs [13]. The Lau et al., study considered four populations with 1.9–2.3 million SNPs per population, compared with ~7.2 million SNPs in the single population considered in the present study. The increased map length with addition of SNPs, which was also observed over both HapMap phase releases, is likely linked to improved resolution of LD structure in previously poorly covered regions, as suggested by Pengelly et al. [6]. The latter study, using WGS data for 96 individuals from the CEU population determined an LDU length of ~1021 for chromosome 22. This compares with ~1184 LDUs from the present study (Table 1), however, a MAF cut-off of 0.05 was used in the earlier study (unlike 0.01 used here), which would contribute to the difference in map length.

The chromosome average kb/LDU ratio (Table 2) suggests that LD extends across the autosomes for ~42 kb in this population, somewhat less extensive than earlier estimates from incompletely saturated maps using tag SNP array data [11, 13]. Comparison with the genetic linkage map lengths of chromosomes in centimorgans [20] (Table 1, Supplementary Fig. 3) confirms the strong correlation ($R^2 = 0.985$) between LDU and recombination map lengths, which must reflect a high degree of positional alignment between historical and present day recombination events [11]. However, on finer-scales the correlation deteriorates in part because of the much lower resolution of genetic linkage maps and the influence of other processes, including selection, mutation and drift, which impact the LD structure. The present maps indicate an average of ~18.4 LDU/cM with a range of 16.7–19.9 for individual chromosomes (Supplementary Fig. 3, Table 1).

Zhang et al. [12] estimated an 'effective population bottleneck time' of 43,000 years based on an estimate of 59,000 LDUs for an autosomal euchromatic genome spanning 34.36 Morgans. The current data suggest 63,428 LDUs/34.36 Morgans (Table 1) = 1846 generations or 46,150 years since an effective bottleneck, assuming 25 years per generation. Consistent with the previous study, the effective bottleneck time reflects the compound effect of numerous population bottlenecks and not any single 'out of Africa' event.

Extent of LD in genic, exonic, intronic and intergenic regions

We compared 16,742 genic regions with 16,720 intergenic regions (Supplementary Table 3). Genic regions (introns and exons combined) comprise ~40% of the sequence, intergenic regions ~55% and centromeric regions ~4.3% (Supplementary Table 1). Comparable LDU lengths (Supplementary Table 2) are ~38, ~61 and ~0.32% the greatly reduced LDU lengths in centromeric regions reflecting deeply suppressed recombination and therefore particularly strong LD. The extent of LD in centromeric regions (Table 2) is dramatically different from the chromosome average of ~42 kb being in the range 140 kb (chromosome 19) to 3773 kb (chromosome 1). The average extent of LD across the genic regions of autosomes is ~44.5 kb compared with ~37.8 kb for intergenic regions (Table 2). Hence, LD is ~16% more extensive in genic compared with intergenic regions presumably reflecting relatively reduced recombination and/or increased selection across genic regions.

We determined the LDU lengths of individual gene exons and introns but quantified these for all 18,268 genes, excluding overlaps. The former span ~2.23% of the genome sequence length and the latter span ~35.53% (Supplementary Table 2). The overall difference in the extent of LD between exons and introns is small (~4%, Table 2) with more extensive LD in exons, however, there is a consistent difference across chromosomes (Table 2) with the difference approaching significance ($P = 0.078$, Table 3). The greater extent of LD across exonic and intronic regions compared with intergenic regions is highly significant ($P < 0.001$, Table 3). The strong relationship between the extent of LD and chromosome recombination rate (Fig. 1) is evident with elevated recombination rates across the smaller chromosomes (e.g., cM/Mb) reflected in markedly reduced extent of LD for these chromosomes. We compared the extent of LD across non-coding RNAs (ncRNA), which comprise ~11% of the sequence (Supplementary Table 2). The extent of LD across these regions is not significantly different from the extent in exons (Table 3) and LD is ~4.5% more extensive than in intronic regions and ~21% more extensive than in intergenic regions (Table 3).

Variable extent of LD across genes from 5′ to 3′ ends

The profile (Fig. 2) shows more extensive LD in the exonic, compared with intronic regions (bins 2–5, but the difference

Table 3 Comparisons of the extent of linkage disequilibrium in kilobases

Variable1 (V1)	Variable 2 (V2)	Chromosome mean V1 (kb)[a]	Chromosome mean V2 (kb)[a]	Difference in extent of LD (kb)/% difference	*P*-value[b]
Genic regions	Intergenic regions	44.54	37.78	6.76/16.4	<0.001
Exons	Introns	46.39	44.49	1.89/4.2	0.078
Exons	Non-coding RNAs	46.39	46.52	0.13/0.3	0.469
Exons	Intergenic regions	46.39	37.78	8.61/20.5	<0.001
Introns	Non-coding RNAs	44.49	46.52	2.02/4.5	0.005
Introns	Intergenic regions	44.49	37.78	6.71/16.3	<0.001
Non-coding RNAs	Intergenic regions	46.52	37.78	8.74/20.7	<0.001

[a]Pairwise comparison of extent of LD in different genome regions in kilobases

[b]*P*-value for differences in extent of LD across all autosomal chromosomes, paired *T*-test (21 degrees of freedom:22 autosomes)

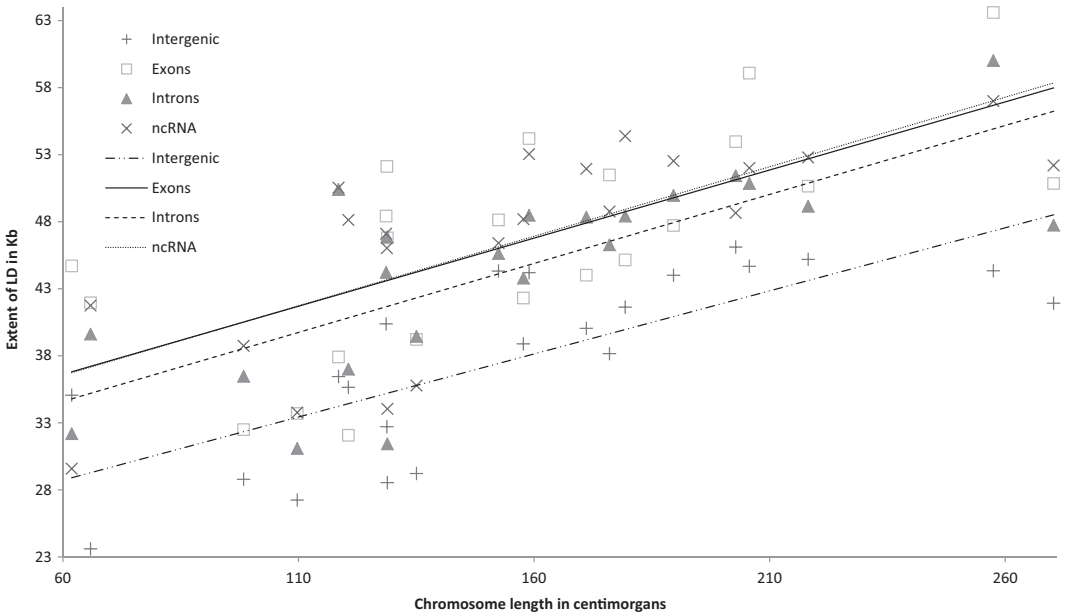

Fig. 1 The extent of LD (in kb) for chromosomes 1–22 against chromosome length in centimorgans. There is a strong linear relationship between chromosome genetic length and extent of LD because smaller chromosomes have a higher rate of recombination per unit physical length. Intergenic regions show significantly reduced extent of LD, intronic regions occupying an intermediate position between exonic and ncRNA regions, which show the most extensive LD. The observed patterns indicate elevated selection and/or reduced recombination in functionally sensitive genome regions

is only significant for bin 2, $P = 0.011$, Supplementary Table 4). In contrast, in bin 1, which is closest to the 5′ end of genes, there is significant evidence that introns have more extensive LD than exons ($P = 0.009$, Supplementary Table 4). Reduced LD in exons towards the 5′ end of genes aligns with the recombination patterns reported by Kong et al. [20] who found bins containing the first exon of a gene have a higher recombination rate than those containing the last exon.

Exons within the more central regions show elevated LD extending to ~52 kb for bin 2 (Supplementary Table 4) with a decline in extent of LD towards the 3′ end. Introns show more uniform LD patterns across the gene with a decline in extent towards the 3′ end. More extensive LD among introns at the 5′ end compared with the 3′ end of genes might reflect increased conservation of first introns, which are enriched for active transcriptional signals, which are under increased selection [22]. The first introns and exons of genes are noted to be more GC rich than the last and internal [23, 24] a feature, which may be related to regulatory functions. Correlations between regions of high GC content and recombination are well established [25] and differential GC content/recombination across genes, might account for some of the variability in extent of LD shown in Fig. 2. However, the much lower resolution of recombination maps makes evaluation of

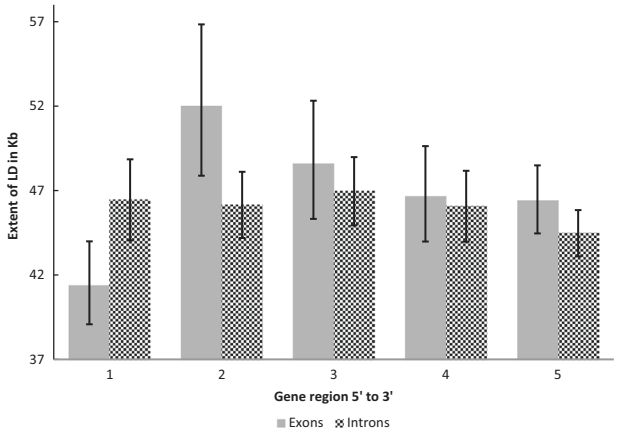

Fig. 2 The extent of LD in kb across the gene profile from 5′ to 3′. The profile of LD across all genes with LDU and kb data allocated to one of five (equally sized within a gene) positional bins. The extent of LD is most variable for exonic regions: LD is less extensive, suggesting relatively increased recombination and/or reduced selection, towards the 5′ end of genes. The 95% confidence intervals are shown

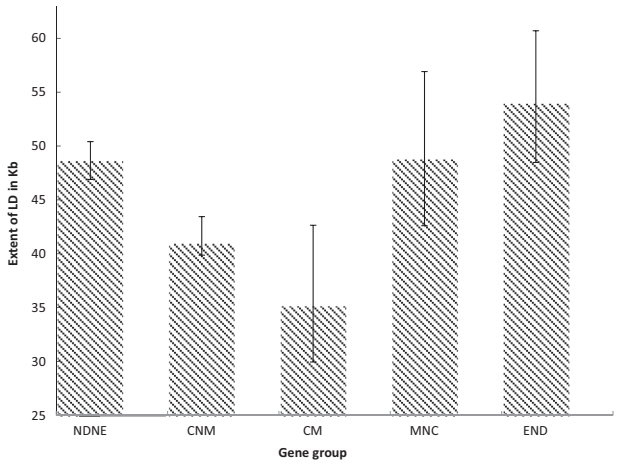

Fig. 3 The extent of LD in kb for different gene groups. Groups of genes classified according to essentiality and relationship to disease phenotypes show wide variability in the extent of LD. Genes classed as essential (END) and Mendelian disease genes (MNC) show elevated extent of LD compared with genes with variation related to complex disease phenotypes (CNM, CM). This might reflect elevated selective pressure within genes assigned to the Mendelian and essential groups. The large group of genes not known to be associated with disease and not known to be essential (NDNE) also show extensive LD but this group is likely to include some miss-classified genes not currently identified as essential or disease related. The 95% confidence intervals are shown

recombination rates for very fine-scale genomic features challenging.

Considering genes stratified into small and large size groups (Supplementary Tables 5 and 6, Supplementary Fig. 4) there is no significant difference in the extent of LD between exons and introns of small genes but increased evidence for a difference in some bins for large genes. LD also extends further generally for large genes in both exonic and intronic regions compared with small genes. While it is possible that larger genes are subject to elevated selective pressure since they have a higher density of exons corresponding to multiple linked sites [26] further studies are required to fully interpret this difference.

Variable extent of LD across gene groups

Essential non-disease (END) genes (Fig. 3, Supplementary Table 7) show significantly more extensive LD (53.9 kb) compared with genes implicated in complex phenotypes (CNM and CM groups, 40.9 and 35.2 kb, respectively). However, the small increase in extent of LD relative to Mendelian genes (MNC) is not significant. Increased selective pressure in both END and MNC gene groups, relative to genes involved in complex phenotypes where variants have reduced phenotypic effect, might account for this difference. The large group of genes classed as non-disease and non-essential (NDNE) also show extensive LD although the extent to which some genes in this group are miss-classified because relationships with disease phenotypes and essentiality are not yet known is unclear.

Discussion

The broad characteristics of LD maps constructed from WGS data compare quite closely with the previously constructed SNP array-based maps. Map lengths are of similar magnitudes, despite vastly more SNPs in the WGS maps, and the computed effective bottleneck time of ~46,000 years is comparable with the previous estimate from a similar population. The pattern of extensive LD across centromeres (shown as broad 'plateau' regions in Supplementary Fig. 2), as previously noted but not directly quantified, suggests that LD extends ~one megabase on average in these regions. The high-resolution maps demonstrate that differences in fine-scale LD structure are detectable down to at least exon level, despite exons encompassing only ~2% of the autosomal genome with an average exon size of just ~300 base pairs. The resolution of this map contrasts, for example, with the Kong et al. [16] study, which examined recombination patterns across a linkage map with ~10-kb resolution.

Berger et al. [17] did not observe less extensive LD on the smaller chromosomes compared with larger chromosomes, however, substantial differences are evident in the maps presented here reflecting the much increased recombination rate of smaller chromosomes. For example, LD extends on average only ~30 kb on chromosome 22 compared with ~50 kb on chromosome 1 (Table 2). The difference in recombination rate mirrors this observation [20] since chromosome 22 recombines at a rate of ~2.1 cM/Mb compared with ~1.1 for chromosome 1 and confirms the

close alignment between extent of LD and recombination rate. However, it is worth noting that, among other differences between the two studies, Berger et al. [17] used the r^2 metric to define pairwise LD.

Despite the differences in methodology, Berger et al. demonstrated more extensive LD (an increase of 13.6%), compared with non-genic regions, which compares quite closely with the findings from this study (~16% more extensive LD in gene regions, Table 3). Genic regions are more conserved than non-genic regions and therefore the tendency for recombination to be higher away from genic regions might reduce disruption of biological pathways [17]. Hence chromosome regions with low recombination are generally found to be enriched for highly conserved genes with essential cellular functions [7, 27]. However, genomic regions or genes with low recombination rates may also have an excess of damaging variation, through higher proportions of rare (MAF <0.01) and non-synonymous variants, because purifying selection is less effective given low recombination [27]. Medically relevant rare variants are more likely to be found in recombination cold-spots: ultra-sensitive regions of the genome, which have up to 400X enrichment of disease-causing variation [28] have a greatly elevated number of recombination cold-spots. The more extensive LD in exonic regions is considered to reflect selection against recombination within exons, and the possibility that recombination is mutagenic [14, 29]. Brick et al. [30] indicate that the *PRDM9* mechanism re-routes meiotic double-stranded breaks away from important genomic functional elements, which may have a protective role against the possible mutagenic effects of recombination. Therefore, there is a complex relationship between low recombination rates, which may allow the build-up of damaging variation, and high recombination, which may be mutagenic, and the interaction with selection.

The variation seen in the LD profile across the exonic and intronic components of genes (Fig. 2) shows some interesting alignment with other studies, which have evaluated the exon–intron architecture of genes [31]. The authors examined correlations with exon and intron ordinal positions in genes for 13 species and found reductions in exon and intron length, GC content and nucleotide divergence with increasing ordinal position from 5′ to 3′. While the functional basis for the patterns of variation are not well understood, the authors argue the relationships observed might reflect time-sequential evolution (earlier arising introns and exons being longer or more divergent).

Discriminating between components of the LD structure determined by recombination from those which reflect positive selection remains extremely challenging. This difficulty extends even to the genetic region for lactase persistence (a region widely recognised as showing strong positive selection), which is associated with at least five regulatory SNPs in a 14-kb region upstream of the *LCT* gene [32]. The authors determine that, although extended haplotype homozygosity (EHH) analysis shows extended haplotype lengths around the selected alleles (consistent with positive selection) an ancestral haplotype also has an extended length for reasons, which are not likely to be related to selection for lactase persistence. The region of EHH aligns precisely with a region with reduced recombination and therefore strong LD in all populations. Regions of the genome with restricted recombination can therefore provide misleading interpretations of the extent of selection.

If the similar patterns of strong LD observed here for exonic and ncRNA regions (Fig. 2) reflect increased positive selection this might align with evidence for the functional significance of ncRNA regions. LD is 4.5% and 20.7% more extensive in ncRNA regions than intronic and intergenic sequences, respectively (Table 3).The ENCODE Project Consortium [33] indicated that there are thousands of ncRNAs and the genome is "pervasively transcribed", suggesting they may have important functions. The ncRNA set includes long non-coding RNAs (RNA transcripts with length >200), which undergo splicing as mRNA precursors, and have been implicated in many significant biological phenomena [34] including imprinting, chromosome conformation, regulation of enzymatic activity, coordination of cell state, differentiation, development and disease. Interestingly, organismal complexity is more closely related to the diversity and size of non-coding RNA expression profiles than with that of protein-coding genes [34]. Further analysis of LD structure differences between different ncRNA classes might be indicative of the relative functional importance of the subtypes.

The findings demonstrate that LD structure provides insights into genome function at the subgenic level with demonstrable differences between LD patterns within features as small as exons. Furthermore, the pattern of LD varies across the gene profile although the functional implications of this are not fully understood. Further analysis of fine-scale LD structure in more genome sequence samples are likely to provide further insights into the functional significance of these patterns.

Data availability

The LDU maps constructed and described in this study are available at: https://doi.org/10.6084/m9.figshare.7850882.v1.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002;296:2225–9.
2. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet. 2006;38:556–60.
3. Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, Pedersen NL, et al. The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet. 2008;83:787–94.
4. Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 2001;29:217–22.
5. Myers S, Freeman C, Auton A, Donnelly P, McVean G. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet. 2008;40:1124–9.
6. Pengelly RJ, Tapper W, Gibson J, Knut M, Tearle R, Collins A, et al. Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. BMC Genom. 2015;16:666.
7. Gibson J, Tapper W, Ennis S, Collins A. Exome-based linkage disequilibrium maps of individual genes: functional clustering and relationship to disease. Hum Genet. 2013;132:233–43.
8. Pengelly RJ, Vergara-Lope A, Alyousfi D, Jabalameli MR, Collins A. Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation. Brief Bioinform. 2017;20:267–73. bbx110.
9. Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, et al. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci USA. 2002;99:2228–33.
10. Zhang W, Collins A, Maniatis N, Tapper W, Morton NE. Properties of linkage disequilibrium (LD) maps. Proc Natl Acad Sci USA. 2002;99:17004–7.
11. Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE. A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci USA. 2005;102:11835–9.
12. Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, et al. Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. Proc Natl Acad Sci USA. 2004;101:18075–80.
13. Lau W, Kuo TY, Tapper W, Cox S, Collins A. Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. Bioinformatics. 2006;23:517–9.
14. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304:581–4.
15. Eberle MA, Rieder MJ, Kruglyak L, Nickerson DA. Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genet. 2006;2:e142.
16. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467:1099–103.
17. Berger S, Schlather M, de Los Campos G, Weigend S, Preisinger R, Erbe M, et al. A scale-corrected comparison of linkage disequilibrium levels between genic and non-genic regions. PLoS ONE. 2015;10:e0141216.
18. Erikson GA, Bodian DL, Rueda M, Molparia B, Scott ER, Scott-Van Zeeland AA, et al. Whole-genome sequencing of a healthy aging cohort. Cell. 2016;165:1002–11.
19. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76:887–93.
20. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002;31:241–7.
21. Spataro N, Rodríguez JA, Navarro A, Bosch E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. Hum Mol Genet. 2017;26:489–500.
22. Park SG, Hannenhalli S, Choi SS. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genom. 2014;15:526.
23. Kalari KR, Casavant M, Bair TB, Keen HL, Comeron JM, Casavant TL, et al. First exons and introns–a survey of GC content and gene structure in the human genome. Silico Biol. 2006;6:237–42.
24. Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. 2012;1:543–556.
25. Fullerton SM, Bernardo Carvalho A, Clark AG. Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol. 2001;18:1139–42.
26. Payseur BA, Nachman MW. Gene density and human nucleotide polymorphism. Mol Biol Evol. 2002;19:336–40.
27. Hussin JG, Hodgkinson A, Idaghdour Y, Grenier JC, Goulet JP, Gbeha E, et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat Genet. 2015;47:400–4.
28. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342:1235587.
29. Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet. 2003;72:1527–35.
30. Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. Genetic recombination is directed away from functional genomic elements in mice. Nature. 2012;485:642.
31. Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genom. 2009;10:47.
32. Liebert A, López S, Jones BL, Montalva N, Gerbault P, Lau W, et al. World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. Hum Genet. 2017;136:1445–53.
33. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447:799.
34. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet. 2016;17:47.