​​In this section we consider the related aspect of haplotype definition in practice. Generally, it is now straight-forward to call SNPs or indel variants. It is however non-trivial to connect individual variants to their haplotype origins. For that reason, algorithms have been developed for phasing, imputation and genealogy inference. It should be noted that these are all different facets of the same problem, and all rely on the foundation of haplotype structures. We discuss what software are close to our definition. In practice, three main paradigms are commonly in use to operationally define haplotypes, all revolving around the core principle of recombined segments: windows, segments and trees. We outline existing methods in the box (Box 3).Th third family of methods appears the closest to this definition. While we are emphasizing branch and genealogy configuration, we would like discus methods, that consider branches below.

In order to be able to use ARG to answr question on first facs a problem of th infering branches. This can be currently achieved in several different ways: inference of full ARG, (<https://academic.oup.com/bioinformatics/article/22/6/768/296494?login=true>, <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-40> , ARG-weaver, argweaver-D), inference of tree topologies (tsinfer, tsdate, Relate).

Genealogy-aware algorithms tsinfer and Relate, use simplified representation of the ARG: succinct tree sequence. Tree sequence format (cite tskit) is a collection of the tables, which contains information of  genealogy of the sample by tracking ancestral and derived nodes, mutations and their location and, most importantly, it contains information of the points where genealogy is switched by recombination (edges in their terminology). Therefore, tree sequence format in practice is tailored to very efficintly store all possible information about haplotype block in it’s branch definition. However, the infrenc algorithm doesn’t allow for overlapping blocks, but rather segments or tracks (Box2: Li, Stephens). In tsinfer ancestral haplotype reconstruction serves as a basis for block length definition and uses HMM later on. Advantag of Relate, that branches are dated. Notably, tsinfer does not explicitly infer an ARG but rather a sequence of local gene trees, described by their topologies only. Like RENT+ and tsinfer, RELATE produces a point estimate of the genealogy, with no allowance for gene tree uncertainty, a well-known problem in gene tree reconciliation coalescent-based methods. It is worth noting that RELATE and tsinfer have one major similarity in that they both use approaches based on a Li-and-Stephens algorithm [111] to estimate ancestral relatedness.

The state space of every possible ancestral history of a sample of genomes is astronomically large and intractable. However, ARGweaver tackles this issue by discretizing time, such that recombination and coalescence events can only occur in a set of discrete time points rendering the ARG space finite. ARGweaver uses MCMC to sample ARGs from a posterior distribution under the SMC. From full conditional probability distribution of all possible ARGs, it then samples the history of a single haplotype using a modified Gibb’s sampler. Although it is the only available method that samples from the posterior distribution of ARGs up to a genome scale, it is computationally expensive and cannot be applied to more than ~50 samples at a time. Moreover, ARGWeaver tends to be sensitive towards  user-defined parameters such as mutation and recombination rate, number of sampling iterations

Using ARG-based methods opens up new opportunities for inference. Haplotype and tree-based inference was demonstrated to be efficient in multiple applications. For example Relate allows to test for selection, estimate effective population size, estimate mutation rates. “The integration of genealogical relationships with genomic variation data has value beyond population and personal genetics, for example in potentially correcting for the differential geographical confounding of rare and common variants in genetic association, and, as Speidel et al.[42](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6726478/#R42) have shown, enables powerful inference of underlying evolutionary events, processes and parameters, such as mutation age, natural selection and ancient contacts between populations.”

Before we identified deficiency in access to overlapping block structure. While not considering ARG, some methods allow for overlapping blocks (HaploBlocker). HaploBlocker then selects? Results in selection and up.

Going back to practical application of ARG definition and methods related to it w face the qustions of vailability of the methods described above depend on sequencing technology. With th new arising tchnologis it will soon bcom possible and routine?

# Drafts

Therefore, tree sequence format in practice is tailored to store all possible information about haplotype block in it’s branch definition. While used within a coalescent simulation framework (msprime) it directly stores simulation output, which contains blocks information. However, transferring actual sequence data into tree sequence format requires using algorithms and underlying assumptions we discussed above. In tsinfer ancestral haplotype reconstruction serves as a basis for block length definition and uses HMM later on.

Having solved problems with creation of the data structure itself, this rich data format can be used for a broad range of applications. For example Relate allows to test for selection, estimate effective population size, estimate mutation rates.

tsinfer/tsdate makes it easier to infer genealogies by inferring tree sequences than the full ARG.

* Full ARG: ARGwaver, complex data structure, which is hard to obtain and iterate

The state space of every possible ancestral history of a sample of genomes is astronomically large and intractable. However, ARGweaver tackles this issue by discretizing time, such that recombination and coalescence events can only occur in a set of discrete time points rendering the ARG space finite. ARGweaver uses MCMC to sample ARGs from a posterior distribution under the SMC. From full conditional probability distribution of all possible ARGs, it then samples the history of a single haplotype using a modified Gibb’s sampler. Although it is the only available method that samples from the posterior distribution of ARGs up to a genome scale, it is computationally expensive and cannot be applied to more than ~50 samples at a time. Moreover, ARGWeaver tends to be sensitive towards  user-defined parameters such as mutation and recombination rate, number of sampling iterations

Also mention - argweaver-D ?

Earlier ARG sampling methods -

LMARC - ​​https://academic.oup.com/bioinformatics/article/22/6/768/296494?login=true

ACG - https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-40

Box 3.

*In practice, three main paradigms are commonly in use to operationally define haplotypes, all revolving around the core principle of recombined segments: windows, segments and trees. Often, they can be further divided into simple heuristics to full statistical treatments. We will consider these in turn below. Lastly, there is an entire class of approaches that infer based on the consequence arising from haplotype structures, rather than direct inference of haplotypes per se. We will treat this latter category separately (see Box 2).*]

<FC-edit: save for later> In some cases identification of large block of the genome resembling the haplotype block comes from observing structural rearrengements. Here borders of inversions serve as a natural borders of the haplotype block. For example in sunflower study haploblock regions were defined on the basis of MDS plots (<https://www.nature.com/articles/s41586-020-2467-6>) / [FC: Add Jones et al., 2012, doi: 10.1038/nature10944. Here we have used a model-based clustering approach to define repeatedly adaptive haplotypes, *based on clustering within MDS space* across adjacent windows. But then if we move this to a different section anyway, it doesn’t matter too much.]</FC-edit>

Across the three classes of methods, window-based methods tend to be the simplest. These approaches rely on pre-defined windows of arbitrary length and primarily operate *across* individuals. Ideally, window sizes should be short enough to avoid spanning recombination junctions. In the simplest form, haplotypes are simply operationally defined as the observed allele at the set of segregating sites in the window. Next, summary statistics are calculated per window along the genome. A representative of such an approach is the H12 test to detect selective sweeps {Garud et al., 2015/doi: 10.1371/journal.pgen.1005004}. In this test, haplotypes (as defined above) within a window are rank-ordered by their frequencies, and selective sweeps on new or recurring mutations are associated with a high combined frequency of the two most common haplotypes (H1 and H2). Other applications include ones exploiting local genomic structures (data-driven clustering/DDC in {Jones et al., 2012, doi: 10.1038/nature10944}, see also (Todesco et al., 2020, doi:10.1038/s41586-020-2467-6; Li and Peter, Genetics, doi: 10.1534/genetics.118.301747}). Among these, *Twisst* {Martin - Van Belleghem 2017, doi: 10.1534/genetics.116.194720} … [FC - to be continued]

While simple, window-based methods are straight-forward and can be powerful. In the era of SNP genotyping, it is also practical. [Discuss Clark 1990 algorithm]. [Haplotype blocks as such are pre-defined - dispenses with any complications here…].

One of the popular methods using this approach is haplotype-based statistics used for inferring selection: iHS and family of related methods. Block inference typically starts from the core allele. The core alleles can be either two variants of a single SNP (core SNP) or multiple variants of a region several base pairs long (core region), such as a gene. Next, the genomic region under investigation is extended to a certain SNP, X. Further development (Voight (2006) allowed to integrate EHH over different window sizes. As X defines position of the polymorphism arising in one of the haplotypes, there will be multiple observed length of the block in this case limited by first mutation disrupting the homozygous block. Haplotype frequencies are further used to infer selection.

nSL - Ferrer-Admetlla, et al., *MBE* 2014 - 10.1093/molbev/msu077

Traditional haplotype-based methods

Short intro saying, that majority of the haplotype-based methods don’t consider branches

Refer to box 3

Possibility to use genealogy information

More advanced methods include either tree sequence or full ARG

However, don’t allow for overlapping blocks (Refer to Box2: Li, Stephens) rather segments or tracks

Branch choice as an underlying assumption

Discuss what kind of information we can potentially lose with the methods above

Point out trade-off between computation feasibility and inference of overlapping blocks

Selection methods for Relate

Access to overlapping block structure

Some methods will have access to overlapping blocks (HaploBlocker and new Brownings)

But have no genealogy information

Box 3.  Haplotype-based methods

 The integration of genealogical relationships with genomic variation data has value beyond population and personal genetics, for example in potentially correcting for the differential geographical confounding of rare and common variants in genetic association, and, as Speidel et al.[42](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6726478/#R42) have shown, enables powerful inference of underlying evolutionary events, processes and parameters, such as mutation age, natural selection and ancient contacts between populations. Moreover, our combined analysis of the UKB and TGP data sets demonstrates the potential of also using tree sequences to integrate data sources and, more generally, to build a reference tree sequence structure for human genomic variation that can be updated as new variants are discovered. Such a structure, coupled with efficient algorithms that make use of the tree sequence could enable (and make optimally powerful

It proceeds by reconstructing ancestral genome fragments for each site, and then inferring the relationships among these fragments according to an ancestral copying process. After explaining all haplotypes in this manner, the method outputs a tree sequence. Notably, tsinfer does not explicitly infer an ARG but rather a sequence of local gene trees, described by their topologies only. – no length

Like RENT+ and tsinfer, RELATE produces a point estimate of the genealogy, with no allowance for gene tree uncertainty, a well-known problem in gene tree reconciliation coalescent-based methods

It is worth noting that RELATE and tsinfer have one major similarity in that they both use approaches based on a Li-and-Stephens algorithm [111] to estimate ancestral relatedness

This likelihood-based method uses ARGweaver to efficiently sample coalescent trees from the poste- rior distribution of ARGs at the locus of interest. It then uses a HMM to marginalize out the allele frequency trajectory conditioned on these trees. Because the ARGs sampled by ARGweaver re- flect the assumption of selective neutrality, CLUES treats them as ‘proposals’ only and uses an importance sampling scheme to estimate the selection coefficient. Stern et al. estimated the se- lection coefficients at various pigmentation-associated variants and found evidence of adapta- tion, consistent with previous work.

In principle, ARGs could also aid in making inferences of polygenic selection across a set of loci associated with a trait based on genome-wide association study (GWAS) summary statistics.

**Practical definition of the haplotype block (original structure)**

It is important to note that phasing, imputation, inferring genealogies are all different facets of the same problem, and all rely on the foundation of haplotype structures. In this section we consider how haplotype blocks are defined in practice by different approaches and demonstrate how it is incorporated into different methods of inference. One could distinguish three main paradigms in defining the haplotype block through the inference schemes: considering length of shared segment and using statistical approaches (variations of Markov chain model, including hidden Markov model) to identify transition between IBD vs non-IBD at each position. The third approach is combining genealogy information withstatistical inference. Additionally, some haplotype-based methods use arbitrary window length to define length of the haplotype.

[FC: ​​*In this section we consider the related aspect of haplotype definition in practice. Generally, it is now straight-forward to call SNPs or indel variants. It is however non-trivial to connect individual variants to their haplotype origins. For that reason, algorithms have been developed for phasing, imputation and genealogy inference. It should be noted that these are all different facets of the same problem, and all rely on the foundation of haplotype structures. In practice, three main paradigms are commonly in use to operationally define haplotypes, all revolving around the core principle of recombined segments: windows, segments and trees. Often, they can be further divided into simple heuristics to full statistical treatments. We will consider these in turn below. Lastly, there is an entire class of approaches that infer based on the consequence arising from haplotype structures, rather than direct inference of haplotypes per se. We will treat this latter category separately (see Box 2).*]

Windows

Segments

Trees

* *Segment-based methods*
* *Tree or genealogy-based methods*

* Structural rearrangements

<FC-edit: save for later> In some cases identification of large block of the genome resembling the haplotype block comes from observing structural rearrengements. Here borders of inversions serve as a natural borders of the haplotype block. For example in sunflower study haploblock regions were defined on the basis of MDS plots (<https://www.nature.com/articles/s41586-020-2467-6>) / [FC: Add Jones et al., 2012, doi: 10.1038/nature10944. Here we have used a model-based clustering approach to define repeatedly adaptive haplotypes, *based on clustering within MDS space* across adjacent windows. But then if we move this to a different section anyway, it doesn’t matter too much.]</FC-edit>

* Fixed windows: H1, H2, TWIST
* *Window-based methods*

Across the three classes of methods, window-based methods tend to be the simplest. These approaches rely on pre-defined windows of arbitrary length and primarily operate *across* individuals. Ideally, window sizes should be short enough to avoid spanning recombination junctions. In the simplest form, haplotypes are simply operationally defined as the observed allele at the set of segregating sites in the window. Next, summary statistics are calculated per window along the genome. A representative of such an approach is the H12 test to detect selective sweeps {Garud et al., 2015/doi: 10.1371/journal.pgen.1005004}. In this test, haplotypes (as defined above) within a window are rank-ordered by their frequencies, and selective sweeps on new or recurring mutations are associated with a high combined frequency of the two most common haplotypes (H1 and H2). Other applications include ones exploiting local genomic structures (data-driven clustering/DDC in {Jones et al., 2012, doi: 10.1038/nature10944}, see also (Todesco et al., 2020, doi:10.1038/s41586-020-2467-6; Li and Peter, Genetics, doi: 10.1534/genetics.118.301747}). Among these, *Twisst* {Martin - Van Belleghem 2017, doi: 10.1534/genetics.116.194720} … [FC - to be continued]

While simple, window-based methods are straight-forward and can be powerful. In the era of SNP genotyping, it is also practical. [Discuss Clark 1990 algorithm]. [Haplotype blocks as such are pre-defined - dispenses with any complications here…].

In the next family of methods haplotype statistics (ex. frequency) are calculated within a window of arbitrary length, ideally its length needs to be proportional to previously inferred recombination rate. In the family of methods for detecting selection (H1, H2 statistics and derivatives) will calculate frequency of the haplotype in the window, ignoring information about it’s possible extent. H1 measures the frequency of the most abundant haplotype and H2 measures the frequency of the second most common haplotype. While still providing a reliable signal to detect selection, the disadvantage of this approach is that real haplotype block length is not taken into account, which makes it hard to extend this approach beyond detecting selection. Windows showing outlier values can further be connected into blocks. Here only frequency of the haplotypes are taken into account and a certain number of mutations (distance threshold) allowed to distinguish between different haplotypes.

While window-based methods do not explicitly infer or use information of haplotype block length, they can take into account genealogy of the sample. Some of the examples come from the field of demographic inference and used for quantification of admixture. ABBA-BABA statistics (D-statistics)(cite softwares) commonly used to assess evidence of gene flow between populations or closely related species. Basic of the tree is tree topology and it is designed to look at the branches in a simplified form. gIMble (Lohse, Barton) provides even more comprehensive model, encoding full genealogy of the sample in a form of generating function. Here IM models are explicitly specified in the likelihood function and allow to jointly infer divergence time and effective migration rate.

* Length of the shared segment.

This approach is implemented on runs of homozygosity and LD blocks. Some softwares consider only blocks identical by state (=runs of homozygosity), while others use various inference schemes trying to extend IBS blocks.

* Selection: EHH, iHS, w

One of the popular methods using this approach is haplotype-based statistics used for inferring selection: iHS and family of related methods. Block inference typically starts from the core allele. The core alleles can be either two variants of a single SNP (core SNP) or multiple variants of a region several base pairs long (core region), such as a gene. Next, the genomic region under investigation is extended to a certain SNP, X. Further development (Voight (2006) allowed to integrate EHH over different window sizes. As X defines position of the polymorphism arising in one of the haplotypes, there will be multiple observed length of the block in this case limited by first mutation disrupting the homozygous block. Haplotype frequencies are further used to infer selection.

nSL - Ferrer-Admetlla, et al., *MBE* 2014 - 10.1093/molbev/msu077

* Haplotype (IBD) block identification: hapIBD

* Correlation structure

* Methods based on Markov chain model (MCMC, HMM)

Li and Stephens (2003) proposed a hidden Markov model (HMM) framework with realistic biological assumptions that underpins a large proportion of haplotype-based inference methods. Originally developed to model patterns of linkage disequilibrium, it has since been widely applied to develop analytical tools and address empirical problems, such as, phasing and imputation of genomic data (Browning and Browning 2007, 2015; Stephens and Scheet 2005; Marchini et al. 2007; Howie et al. 2009; Li et al. 2010), inference of population structure and  demographic history (Lawson et al. 2012; Steinrücken et al. 2018, 2019, Hellenthal et al. 2014), characterisation of local admixture (Sundquist et al. 2008; Price et al. 2009), inference of local genealogies (Rasmussen et al. 2014; Kelleher et al. 2019; Speidel et al. 2019), and many more.

Sequential Markov coalescent (SMC) is assumed as an approximation, whenever a Hidden Markov Model is used.

* Using in statistical phasing: phasing softwares make underlying assumptions about haplotype structure
* Inference of demographic history: PSMC, SMC++
* Haplotype block identification: HaploBlocker: haplotype block is a sequence of genetic markers that has a predefined minimum frequency - simplified HMM
* Population structure: ADMIXTURE, fineSTRUCTURE
* Selection?
* Genealogy-aware algorithms

tsinfer/tsdate makes it easier to infer genealogies by inferring tree sequences than the full ARG.

Genealogy-aware algorithms tsinfer and Relate, don’t yet work with full ARG, but instead use it in unwrapped form: succinct tree sequence. Tree sequence format (see tskit) is a collection of the tables, which contains information of  genealogy of the sample by tracking ancestral and derived nodes, mutations and their location and, most importantly, it contains information of the points where genealogy is switched by recombination (edges in their terminology). Therefore, tree sequence format in practice is tailored to store all possible information about haplotype block in it’s branch definition. While used within a coalescent simulation framework (msprime) it directly stores simulation output, which contains blocks information. However, transferring actual sequence data into tree sequence format requires using algorithms and underlying assumptions we discussed above. In tsinfer ancestral haplotype reconstruction serves as a basis for block length definition and uses HMM later on.

Having solved problems with creation of the data structure itself, this rich data format can be used for a broad range of applications. For example Relate allows to test for selection, estimate effective population size, estimate mutation rates

* Full ARG: ARGwaver, complex data structure, which is hard to obtain and iterate

The state space of every possible ancestral history of a sample of genomes is astronomically large and intractable. However, ARGweaver tackles this issue by discretizing time, such that recombination and coalescence events can only occur in a set of discrete time points rendering the ARG space finite. ARGweaver uses MCMC to sample ARGs from a posterior distribution under the SMC. From full conditional probability distribution of all possible ARGs, it then samples the history of a single haplotype using a modified Gibb’s sampler. Although it is the only available method that samples from the posterior distribution of ARGs up to a genome scale, it is computationally expensive and cannot be applied to more than ~50 samples at a time. Moreover, ARGWeaver tends to be sensitive towards  user-defined parameters such as mutation and recombination rate, number of sampling iterations

Also mention - argweaver-D ?

Earlier ARG sampling methods -

LMARC - ​​https://academic.oup.com/bioinformatics/article/22/6/768/296494?login=true

ACG - https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-40

*Tsinfer* is based on an efficient data structure called tree sequence (Kelleher et al. 2016), which makes use of the correlation structure between adjacent trees along the genome. Using a strategy that infers the ancestral haplotypes, as well as the copying paths for them and the sample haplotypes, *tsinfer* achieves extraordinary efficiency to be scaled to hundreds of thousands of samples at the same time (Kelleher et al. 2019). On the other hand, *Relate* is a two-step approach that infers different elements of the genealogies: tree topologies and branch lengths, using position-specific distance matrices and a Markov Chain Monte Carlo (MCMC) algorithm, respectively.