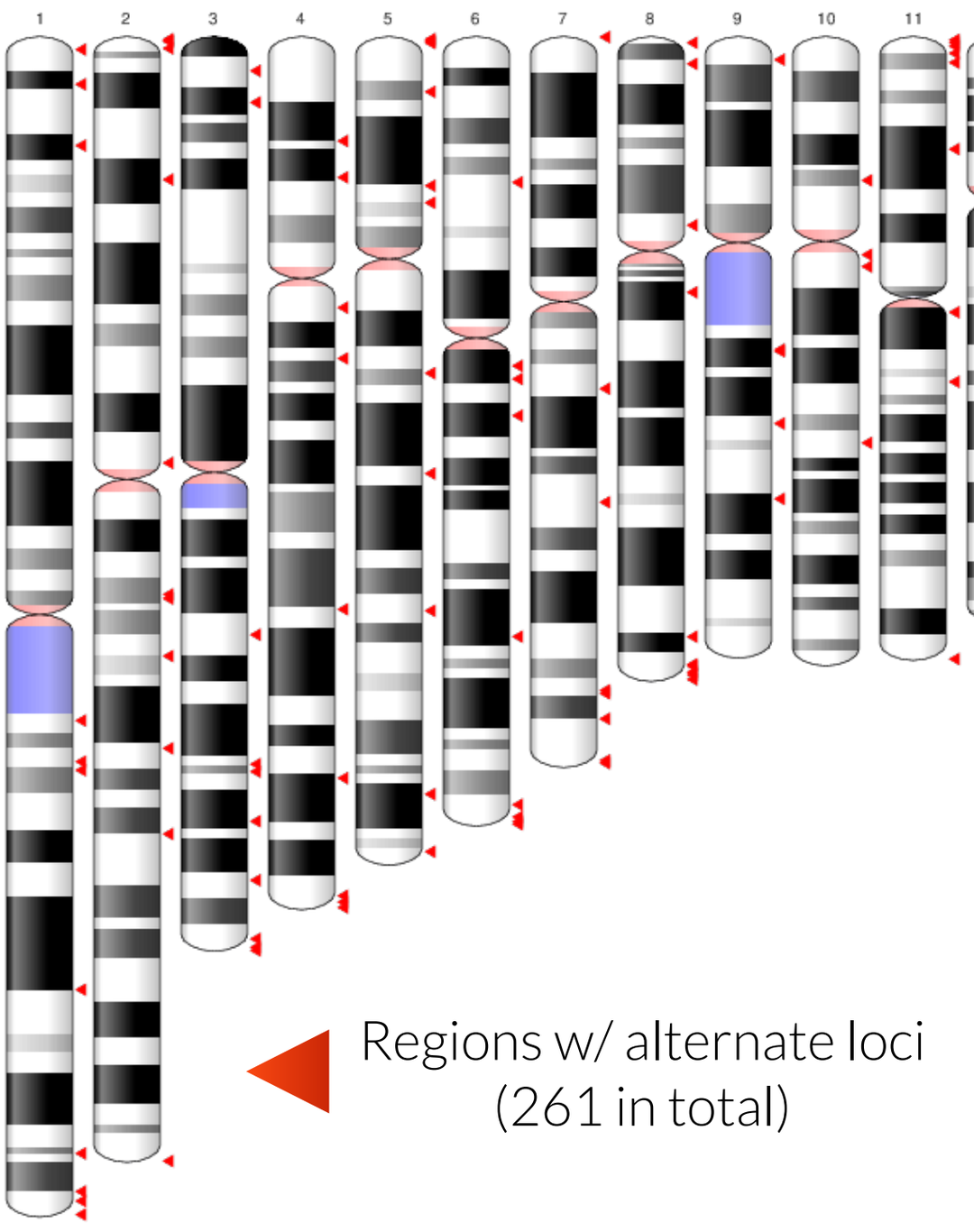
Variant calling while accounting for alternate haplotypes

Genome Reference Consortium Workshop
Sept 21, 2014

Aaron Quinlan
University of Virginia

quinlanlab.org

Motivation: HG has long had alt loci, but we don't handle them properly



Regions w/ alternate loci
(261 in total)

Current assembly model

The Genome Reference Consortium (GRC) introduced a new assembly model with the production of GRCh37. (see PMID: 21750661) This model allows for the representation of alternate sequences for complex genomic regions. These are regions that have high degrees of sequence diversity and for which it would be difficult to define a single path that would capture all diversity. An example of such a region is the MHC region on chr6. Currently, 8 allelic representations of this region are available, the path that is incorporated into the chromosome assembly and 7 alternate paths representing different alleles. Additionally, there is a FIX patch that corrects one of these sequence. Assemblies containing alternate sequences are often referred to as 'complex assemblies'.

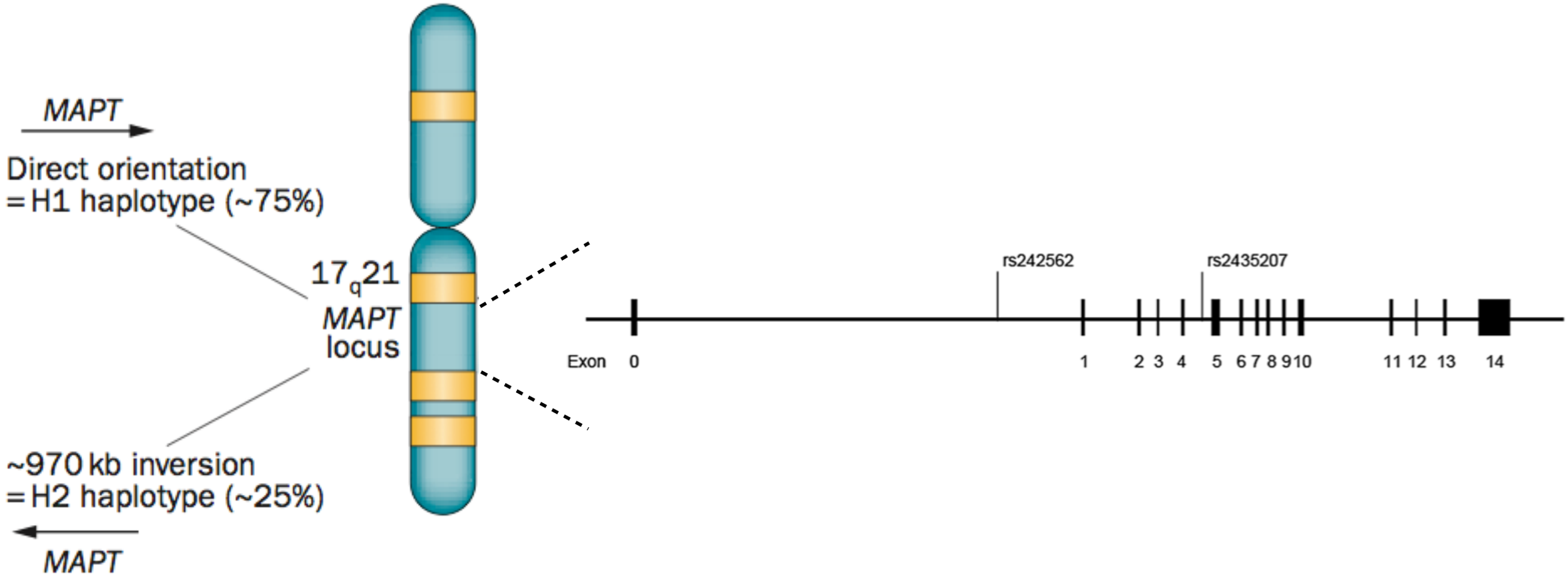
The problem: most tools and file formats expect a single, haploid representation of the assembly. When the alternate sequences are added, it is typically impossible for tools to distinguish between allelic duplication (which should not be penalized in analysis) and duplicated sequence (repeats) in the genome.

The new assembly data model approaches this in two ways. Sequences are partitioned into 'Assembly Units'. The 'Primary Assembly Unit' contains the non-redundant haploid representation of the assembly. Alternate sequences are placed in separate assembly units; as long as the alternate sequences are not from the same region they can be in the same unit. At the MHC, where there are 7 alternate sequences, there are 7 additional Assembly Units (ALT_LOCI_1 - ALT_LOCI_7). Additionally, alternate sequences are aligned to the chromosome. Regions on the chromosome that contain an alignment to an alternate sequence are annotated and named. The alignments of the alternate sequences and the sequences in the Primary assembly are released as part of the assembly definition. Note: for some complex assemblies (like mouse) there are alternate loci that have no alignment to the primary assembly because the allelic diversity is too great and the alternate sequence has not been extended far enough into unique sequence that a good alignment can be generated. These sequences are placed in assembly units other than the primary, but they are not part of any region.

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

Deanna Church and Brad Holmes

A case study: The MAPT locus (17q21.31)



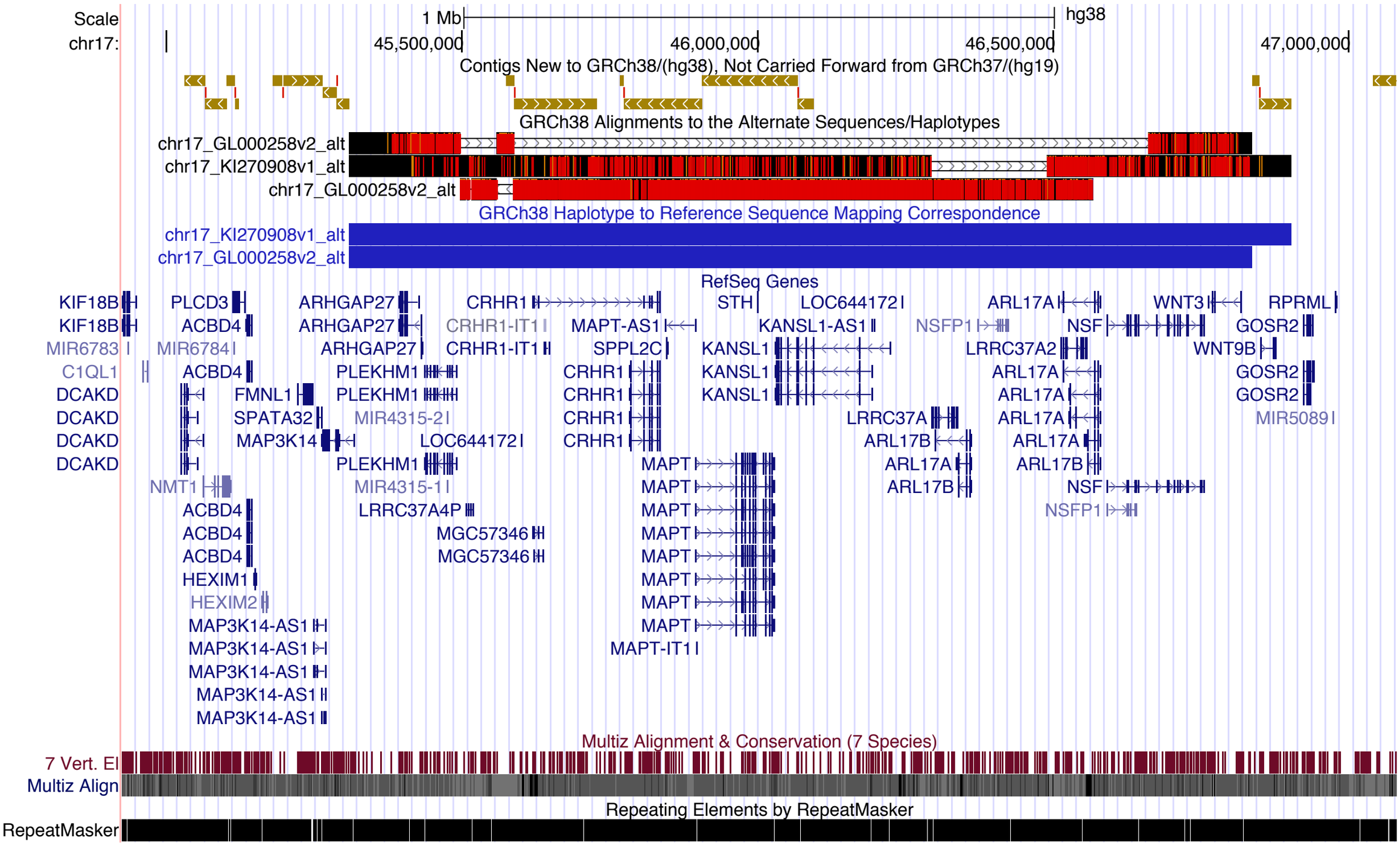
Disease phenotypes associated with each haplotype

H1 associated with Alzheimer and Parkinson diseases

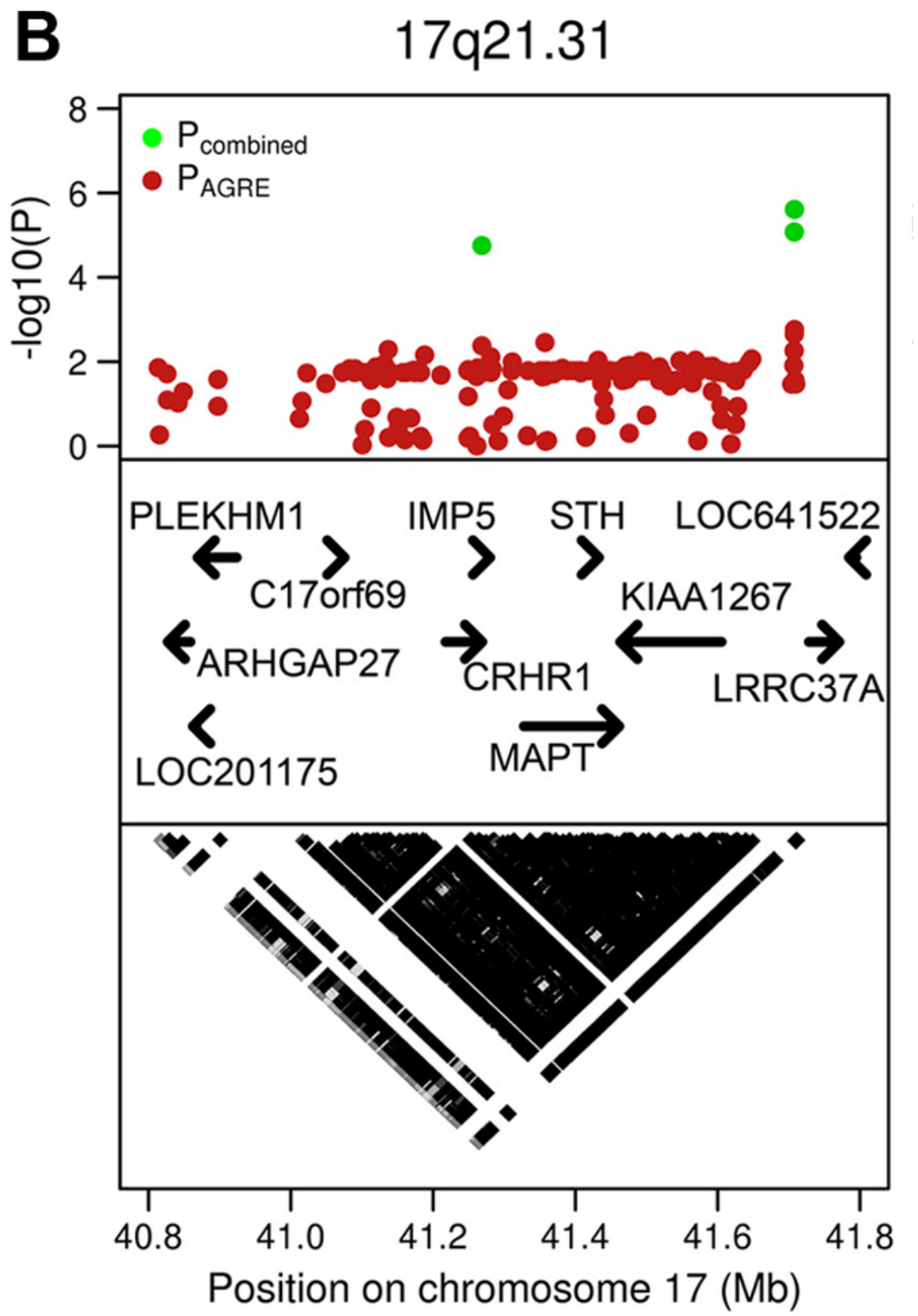
H2 positively selected in Europeans, carriers predisposed to 17q21.31 microdeletion syndrome via ⬆ NAHR

doi:10.1038/nrneurol.2012.169

A case study: The MAPT locus (17q21.31)



H1 and inverted H2 don't recombine



Extensive LD owing to suppressed recombination between H1 and inverted H2

Genetic Analysis of Variation in Human Meiotic Recombination
Reshmi Chowdhury, Philippe R. J. Bois, Eleanor Feingold, Stephanie L. Sherman, Vivian G. Cheung

Several SNPs distiguish H1 from H2

Structural diversity and African origin of the 17q21.31 inversion polymorphism

Karyn Meltz Steinberg[1,11], Francesca Antonacci[1,11], Peter H Sudmant[1], Jeffrey M Kidd[1,10], Catarina D Campbell[1], Laura Vives[1], Maika Malig[1], Laura Scheinfeldt[2], William Beggs[2], Muntaser Ibrahim[3], Godfrey Lema[4], Thomas B Nyambo[4], Sabah A Omar[5], Jean-Marie Bodo[6], Alain Froment[7], Michael P Donnelly[8,10], Kenneth K Kidd[8], Sarah A Tishkoff[2] & Evan E Eichler[1,9]

Supplementary Table 8. Inversion and duplication tagging SNPs

SNP	Genomic position (b36)	Genomic position (b37)	H1'	H2'
rs241039	41070456	43714673	A	T
rs434428	41081467	43725684	G	A
rs241027	41091261	43735478	A	G
rs2049515	41117639	43761856	C	T
rs10491144	41128907	43773124	A	C
rs10514879	41158754	43802971	C	T
rs2902662	41162708	43806925	G	A
rs17563599	41163726	43807955	A	C
rs11079718	41195723	43839951	A	T
rs1396862	41258778	43902997	G	A
rs1078830	41301901	43946112	T	C
rs916793	41310477	43954686	G	A
rs17563986	41347100	43991272	A	G
rs17650901	41395527	44039691	T	C
rs1800547	41407682	44051846	A	G
rs17651213	41407760	44051924	G	A
rs1981997	41412603	44056767	G	A
rs1052553	41429726	44073889	A	G
rs8070723	41436901	44081064	A	G
rs9468	41457408	44101563	T	C

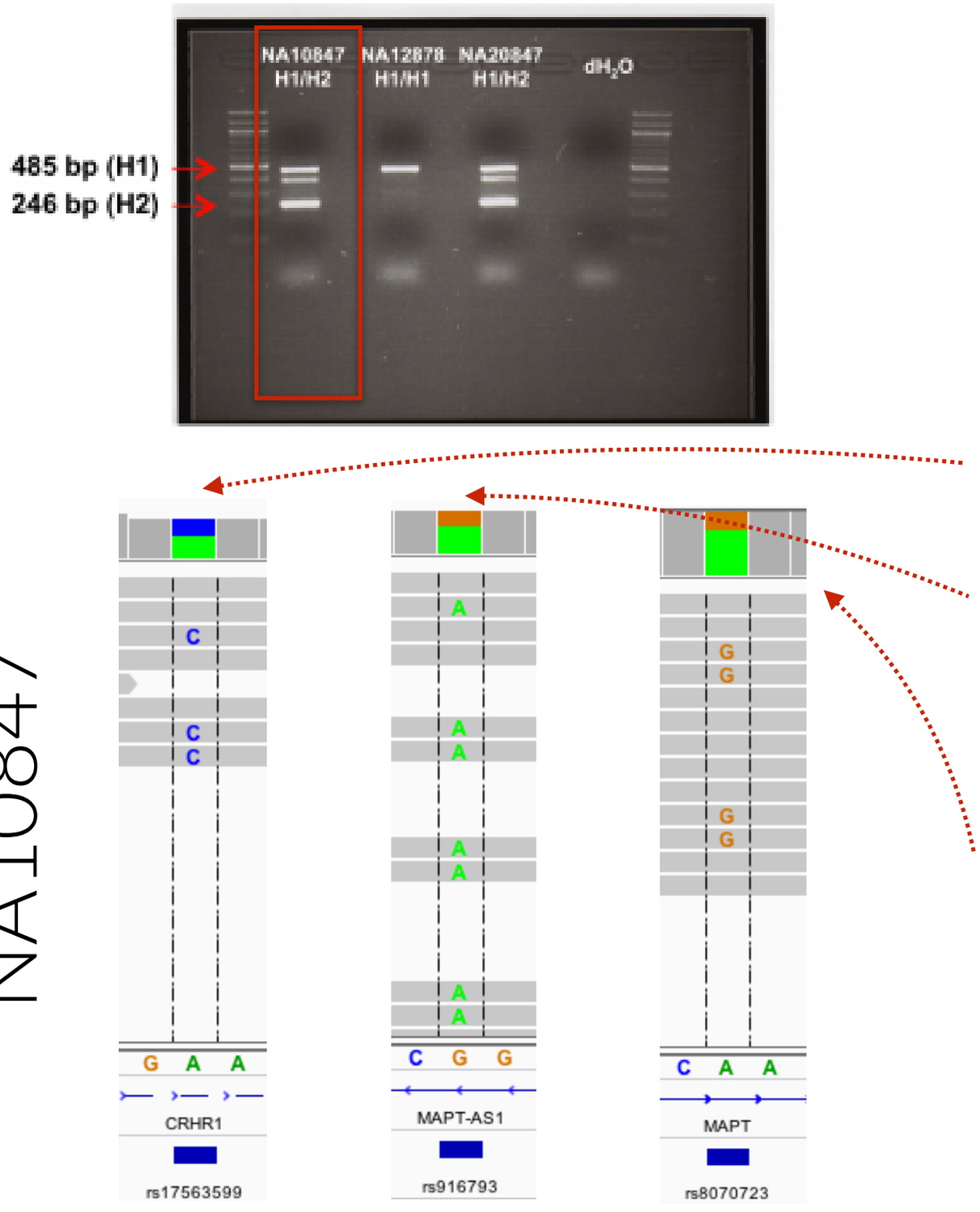
. . .

Kitzman et al: NA10847 is heterozygous H1/H2

Haplotype-resolved genome sequencing of a Gujarati Indian individual

Jacob O Kitzman, Alexandra P MacKenzie, Andrew Adey, Joseph B Hiatt, Rupali P Patwardhan, Peter H Sudmant, Sarah B Ng, Can Alkan, Ruolan Qiu, Evan E Eichler & Jay Shendure

Kitzman et al: NA10847 is heterozygous H1/H2



Haplotype-resolved genome sequencing of a Gujarati Indian individual

Jacob O Kitzman, Alexandra P MacKenzie, Andrew Adey, Joseph B Hiatt, Rupali P Patwardhan, Peter H Sudmant, Sarah B Ng, Can Alkan, Ruolan Qiu, Evan E Eichler & Jay Shendure

Supplementary Table 8. Inversion and duplication tagging SNPs

SNP	Genomic position (b36)	Genomic position (b37)	H1'	H2'
rs241039	41070456	43714673	A	T
rs434428	41081467	43725684	G	A
rs241027	41091261	43735478	A	G
rs2049515	41117639	43761856	C	T
rs10491144	41128907	43773124	A	C
rs10514879	41158754	43802971	C	T
rs2902662	41162708	43806925	G	A
rs17563599	41163726	43807955	A	C
rs11079718	41195723	43839951	A	T
rs1396862	41258778	43902997	G	A
rs1078830	41301901	43946112	T	C
rs916793	41310477	43954686	G	A
rs17563986	41347100	43991272	A	G
rs17650901	41395527	44039691	T	C
rs1800547	41407682	44051846	A	G
rs17651213	41407760	44051924	G	A
rs1981997	41412603	44056767	G	A
rs1052553	41429726	44073889	A	G
rs8070723	41436901	44081064	A	G
rs9468	41457408	44101563	T	C
rs12150447	41483977	44128125	A	C
rs2838	41497167	44141347	A	G
rs1468241	41551932	44196153	A	C
rs1528075	41576231	44220454	T	G
rs1528072	41592502	44236725	C	A
rs2668692	41648797	44293020	C	T
rs2957297	41723989	44368212	A	G
rs199457	42150653	44795469	C	C
rs199456	42153103	44797919	C	C
rs199451	42156968	44801784	G	G
rs199448	42164185	44809001	A	A
rs199533	42184098	44828931	C	C

How do we support variant detection with alternate alleles using the VCF format?

NA10847 is heterozygous H1/H2

H1 (chr17)

A G A C A C

H2 (GL000258v2_alt)

T A G T C T

NA10847 is heterozygous H1/H2

H1 (chr17)

A G A C A C

H2 (GL000258v2_alt)

T A G T C T

A G A C A C
T A G T C T

NA10847 (het. **H1**/**H2**)

NA10847 is heterozygous H1/H2

H1 (chr17)

A G A C A C

A G A C A C

T A G T C T

H2 (GL000258v2_alt)

T A G T C T

A G A C A C

T A G T C T

Requires that we align reads to both loci*

A G A C A C

T A G T C T

NA10847 (het. **H1**/**H2**)

* We need to be able to distinguish multiple alignments arising from ALT loci versus multiple mappings arising from segdups, repetitive elements. Else, MAPQs penalized and/or alignments not reported, depending on the behavior of the aligner.

Heng Li has started a discussion about how best to make BWA alt-aware
Colin Hercus has started a discussion about how best to make Novoalign alt-aware

NA10847 is heterozygous H1/H2

H1 (chr17)

A G A C A C
A G A C A C
T A G T C T

H2 (GL000258v2_alt)

T A G T C T
A G A C A C
T A G T C T

Requires that we align reads to both loci

A G A C A C
T A G T C T

NA10847 (het. **H1**/**H2**)

↓

Variant calling

↓

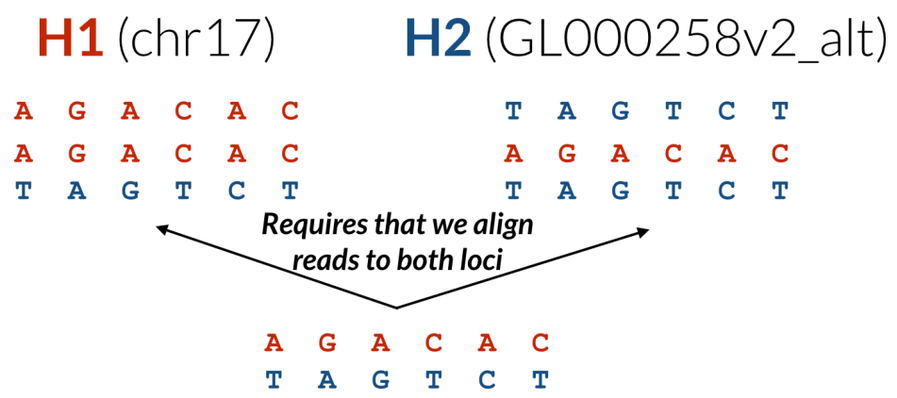
VCF entries for chr17 (H1)

```
chr17  A  T  0/1
chr17  G  A  0/1
chr17  A  G  0/1
chr17  C  T  0/1
chr17  A  C  0/1
chr17  C  T  0/1
```

VCF entries for alt locus (H2)

```
GL000258v2_alt  T  A  0/1
GL000258v2_alt  A  G  0/1
GL000258v2_alt  G  A  0/1
GL000258v2_alt  T  C  0/1
GL000258v2_alt  C  A  0/1
GL000258v2_alt  T  C  0/1
```

NA10847 is heterozygous H1/H2



Note: Allelic relationship is not reflected in "raw" VCF

Variant calling

VCF entries for chr17 (H1)

```
chr17  A  T  0/1
chr17  G  A  0/1
chr17  A  G  0/1
chr17  C  T  0/1
chr17  A  C  0/1
chr17  C  T  0/1
```

VCF entries for alt locus (H2)

```
GL000258v2_alt  T  A  0/1
GL000258v2_alt  A  G  0/1
GL000258v2_alt  G  A  0/1
GL000258v2_alt  T  C  0/1
GL000258v2_alt  C  A  0/1
GL000258v2_alt  T  C  0/1
```

Proposal: Develop a downstream tool that leverages informative SNPs to distinguish and assign haplotypes at alt loci via a standard VCF file.

Intermediate solution until variant callers handle this complexity natively

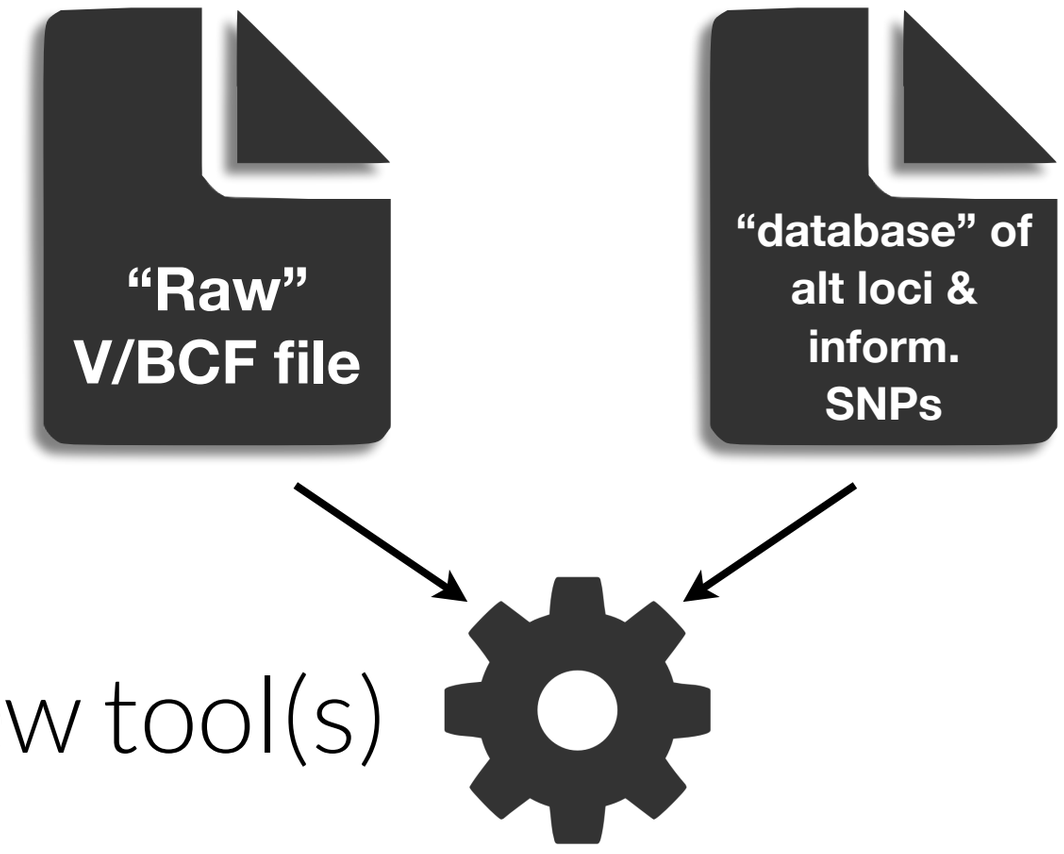
Overview

Overview

alt_locus	main_locus
GL000258.2 (H2)	chr17:45309498-46836265 (H1)

infor. markers	GRCh38 position	H1	H2
rs241039	45637307	A	T
rs2049515	45684490	C	T
…			

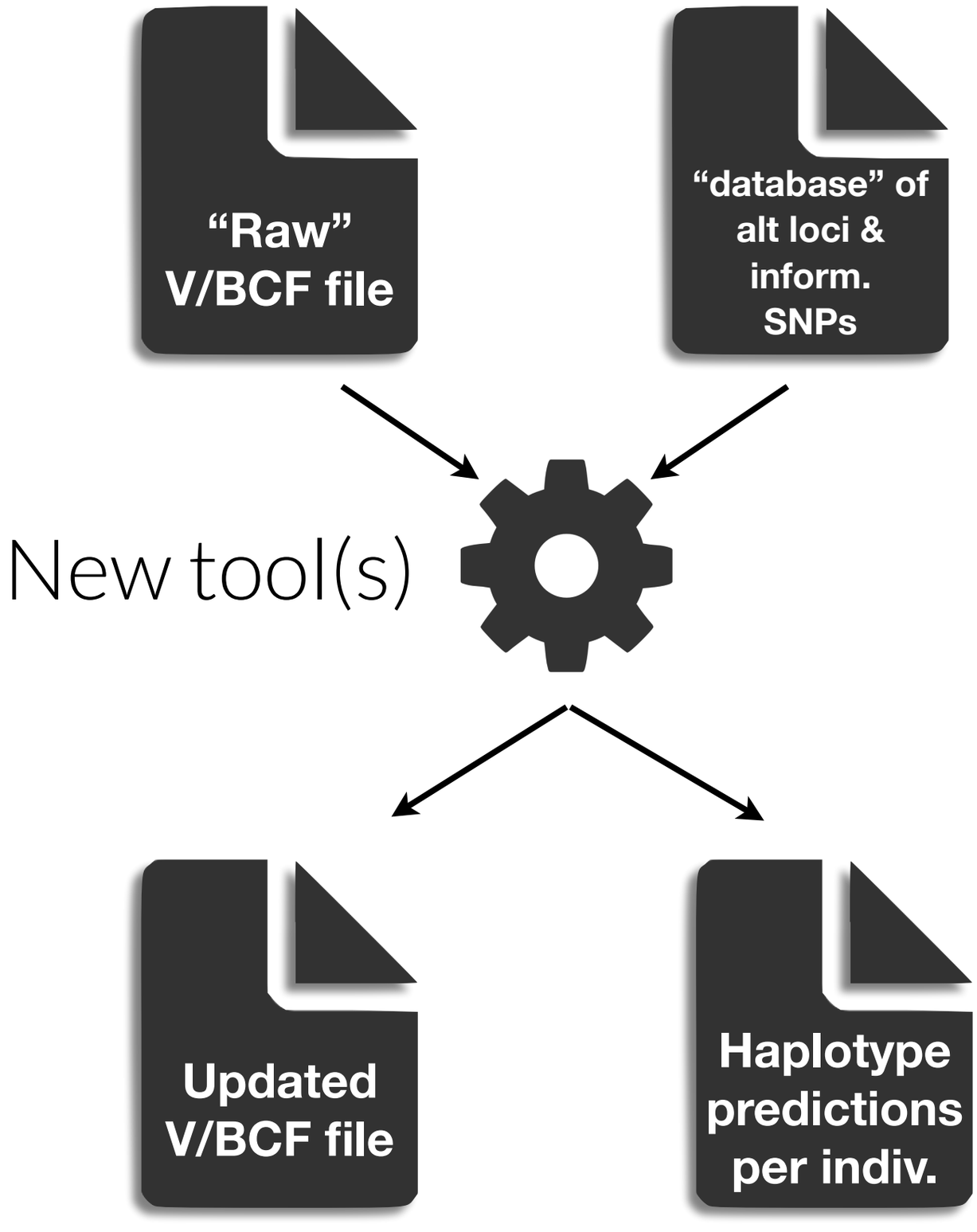
Overview



alt_locus	main_locus
GL000258.2 (H2)	chr17:45309498-46836265 (H1)

infor. markers	GRCh38 position	H1	H2
rs241039	45637307	A	T
rs2049515	45684490	C	T
…			

Overview



alt_locus	main_locus
GL000258.2 (H2)	chr17:45309498-46836265 (H1)

infor. markers	GRCh38 position	H1	H2
rs241039	45637307	A	T
rs2049515	45684490	C	T
…			

Augment VCF with assembly information

```
##seq-info=<name=chr17, id=CM000679.2>

##region-info=<name=MAPT, id=GL000258.2,
assoc_id=CM000679.2, reg=45309498-46836265>
```

Based on ideas from Deanna Church

Introduce new (reserved) VCF INFO tags

```
##INFO=<ID=ALTLOCS,
        Number=.,
        Type=String,
        Description="A list of the alternate
                     loci in the reference
                     genome that are
                     associated with this
                     locus">
```

```
##INFO=<ID=ALTHAPS,
        Number=.,
        Type=String,
        Description="A list of the known
                     haplotypes that are
                     associated with this
                     locus">
```

```
##FORMAT=<ID=HT,Number=1,Type=String,Description="Haplotype
combination based on ALTHAPS">
```

(Draft) example VCF output, post “correction”

```
##seq-info=<name=chr17, id=CM000679.2>
##region-info=<name=MAPT,id=GL000258.2,assoc_id=CM000679.2,reg=45309498-46836265>
##INFO=<ID=ALTLOC,Number=.,Type=String,Description="A list of the alternate loci is
the reference genome that are associated with this locus">
##INFO=<ID=ALTHAP,Number=.,Type=String,Description="A list of the known haplotypes
that are associated with this locus">

#CHROM      POS REF ALT INFO                                    FORMAT      NA10847
chr17       111 A   T   ALTLOCS=GL000258.2;ALTHAPs=H1,H2   GT:HT        0/1:0/1
chr17       222 G   A   ALTLOCS=GL000258.2;ALTHAPs=H1,H2   GT:HT        0/1:0/1
chr17       333 A   G   ALTLOCS=GL000258.2;ALTHAPS=H1,H2   GT:HT        0/1:0/1
. . .
GL000258.2  111 A   T   ALTLOCS=chr17;ALTHAPs=H1,H2        GT:HT        0/1:0/1
GL000258.2  222 G   A   ALTLOCS=chr17;ALTHAPs=H1,H2        GT:HT        0/1:0/1
GL000258.2  333 A   G   ALTLOCS=chr17;ALTHAPS=H1,H2        GT:HT        0/1:0/1
. . .
```

(Draft) example VCF output, post “correction”

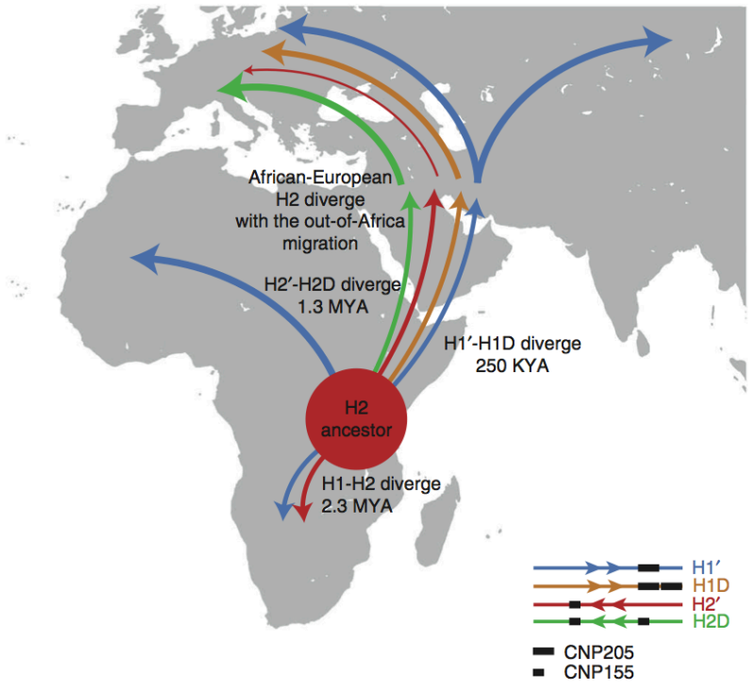
```
##seq-info=<name=chr17, id=CM000679.2>
##region-info=<name=MAPT,id=GL000258.2,assoc_id=CM000679.2,reg=45309498-46836265>
##INFO=<ID=ALTLOC,Number=.,Type=String,Description="A list of the alternate loci is
the reference genome that are associated with this locus">
##INFO=<ID=ALTHAP,Number=.,Type=String,Description="A list of the known haplotypes
that are associated with this locus">

#CHROM      POS REF ALT INFO                                    FORMAT      NA10847
chr17       111 A   T   ALTLOCS=GL000258.2;ALTHAPs=H1,H2        GT:HT       0/1:0/1
chr17       222 G   A   ALTLOCS=GL000258.2;ALTHAPs=H1,H2        GT:HT       0/1:0/1
chr17       333 A   G   ALTLOCS=GL000258.2;ALTHAPS=H1,H2        GT:HT       0/1:0/1
. . .
GL000258.2  111 A   T   ALTLOCS=chr17;ALTHAPs=H1,H2             GT:HT       0/1:0/1
GL000258.2  222 G   A   ALTLOCS=chr17;ALTHAPs=H1,H2             GT:HT       0/1:0/1
GL000258.2  333 A   G   ALTLOCS=chr17;ALTHAPS=H1,H2             GT:HT       0/1:0/1
. . .
```

Solely two different haplotypes is the base case.

For example NA10847 is *actually* H1/H2**D**

Supplementary Table 8. Inversion and duplication tagging SNPs

SNP	Genomic position (b36)	Genomic position (b37)	H1'	H2'	H2D
rs241039	41070456	43714673	A	T	T
rs434428	41081467	43725684	G	A	A
rs241027	41091261	43735478	A	G	G
rs2049515	41117639	43761856	C	T	T
rs10491144	41128907	43773124	A	C	C
rs10514879	41158754	43802971	C	T	T
rs2902662	41162708	43806925	G	A	A
rs17563599	41163726	43807955	A	C	C
rs11079718	41195723	43839951	A	T	T
rs1396862	41258778	43902997	G	A	A
rs1078830	41301901	43946112	T	C	C
rs916793	41310477	43954686	G	A	A
rs17563986	41347100	43991272	A	G	G
rs17650901	41395527	44039691	T	C	C
rs1800547	41407682	44051846	A	G	G
rs17651213	41407760	44051924	G	A	A
rs1981997	41412603	44056767	G	A	A
rs1052553	41429726	44073889	A	G	G
rs8070723	41436901	44081064	A	G	G
rs9468	41457408	44101563	T	C	C
rs12150447	41483977	44128125	A	C	C
rs2838	41497167	44141347	A	G	G
rs1468241	41551932	44196153	A	C	C
rs1528075	41576231	44220454	T	G	G
rs1528072	41592502	44236725	C	A	A
rs2668692	41648797	44293020	C	T	T
rs2957297	41723989	44368212	A	G	G
rs199457	42150653	44795469	C	C	T
rs199456	42153103	44797919	C	C	T
rs199451	42156968	44801784	G	G	A
rs199448	42164185	44809001	A	A	G
rs199533	42184098	44828931	C	C	T



H2D derived from ancestral H2.

Markers that distinguish H2D from H2

NA10847	EUR	CEU			X	X	X					H2D/H1'

Structural diversity and African origin of the 17q21.31 inversion polymorphism

Karyn Meltz Steinberg[1,11], Francesca Antonacci[1,11], Peter H Sudmant[1], Jeffrey M Kidd[1,10], Catarina D Campbell[1], Laura Vives[1], Maika Malig[1], Laura Scheinfeldt[2], William Beggs[2], Muntaser Ibrahim[3], Godfrey Lema[4], Thomas B Nyambo[4], Sabah A Omar[5], Jean-Marie Bodo[6], Alain Froment[7], Michael P Donnelly[8,10], Kenneth K Kidd[8], Sarah A Tishkoff[2] & Evan E Eichler[1,9]

Interpretation is much harder w/ many haplotypes

VCF entries for chr17

```
chr17  A  T  0/1
chr17  G  A  0/1
chr17  A  G  0/1
chr17  C  T  0/1
chr17  A  C  0/1
chr17  C  T  0/1
chr17  C  T  0/1
chr17  C  T  0/1
chr17  G  A  0/1
chr17  A  G  0/1
chr17  C  T  0/1
```

VCF entries for alt locus

```
GL000258v2_alt  T  A  0/1
GL000258v2_alt  A  G  0/1
GL000258v2_alt  G  A  0/1
GL000258v2_alt  T  C  0/1
GL000258v2_alt  C  A  0/1
GL000258v2_alt  T  C  0/1
GL000258v2_alt  C  T  0/1
GL000258v2_alt  C  T  0/1
GL000258v2_alt  G  A  0/1
GL000258v2_alt  A  G  0/1
GL000258v2_alt  C  T  0/1
```

H1 H2 H2D

(Draft) example VCF output, post “correction”

##seq-info=<name=chr17, id=CM000679.2>
##region-info=<name=MAPT,id=GL000258.2,assoc_id=CM000679.2,reg=45309498-46836265>
##INFO=<ID=ALTLOC,Number=.,Type=String,Description=“A list of the alternate loci is the reference genome that are associated with this locus”>
##INFO=<ID=ALTHAP,Number=.,Type=String,Description=“A list of the known haplotypes that are associated with this locus”>

#CHROM	**POS**	**REF**	**ALT**	**INFO**	**FORMAT**	**NA10847**
chr17	111	A	T	**ALTLOCS**=GL000258.2;**ALTHAPs**=H1,H2,H2D	GT:**HT**	0/1:**0/2**
chr17	222	G	A	**ALTLOCS**=GL000258.2;**ALTHAPs**=H1,H2,H2D	GT:**HT**	0/1:**0/2**
chr17	333	A	G	**ALTLOCS**=GL000258.2;**ALTHAPS**=H1,H2,H2D	GT:**HT**	0/1:**0/2**
.	.	.				
GL000258.2	111	A	T	**ALTLOCS**=chr17;**ALTHAPs**=H1,H2,H2D	GT:**HT**	0/1:**0/2**
GL000258.2	222	G	A	**ALTLOCS**=chr17;**ALTHAPs**=H1,H2,H2D	GT:**HT**	0/1:**0/2**
GL000258.2	333	A	G	**ALTLOCS**=chr17;**ALTHAPS**=H1,H2,H2D	GT:**HT**	0/1:**0/2**
.	.	.				

(Draft) example VCF output, post “correction”

```
##seq-info=<name=chr17, id=CM000679.2>
##region-info=<name=MAPT,id=GL000258.2,assoc_id=CM000679.2,reg=45309498-46836265>
##INFO=<ID=ALTLOC,Number=.,Type=String,Description="A list of the alternate loci is the reference genome that are associated with this locus">
##INFO=<ID=ALTHAP,Number=.,Type=String,Description="A list of the known haplotypes that are associated with this locus">

#CHROM      POS REF ALT   INFO                                          FORMAT    NA10847
chr17       111 A   T     ALTLOCS=GL000258.2;ALTHAPs=H1,H2,H2D          GT:HT     0/1:0/2
chr17       222 G   A     ALTLOCS=GL000258.2;ALTHAPs=H1,H2,H2D          GT:HT     0/1:0/2
chr17       333 A   G     ALTLOCS=GL000258.2;ALTHAPS=H1,H2,H2D          GT:HT     0/1:0/2
. . .
GL000258.2  111 A   T     ALTLOCS=chr17;ALTHAPs=H1,H2,H2D               GT:HT     0/1:0/2
GL000258.2  222 G   A     ALTLOCS=chr17;ALTHAPs=H1,H2,H2D               GT:HT     0/1:0/2
GL000258.2  333 A   G     ALTLOCS=chr17;ALTHAPS=H1,H2,H2D               GT:HT     0/1:0/2
. . .
```

sample	chrom	start	end	hap1	hap2	markers
NA10847	chr17	45309498	46836265	**H1**	**H2D**	rs241039,rs434428,…
NA12878	chr17	45309498	46836265	**H1**	**H1**	rs241039,rs434428,…
NA21599	chr17	45309498	46836265	**H2D**	**H2D**	rs241039,rs434428,…

(Draft) example VCF output, post “correction”

```
##seq-info=<name=chr17, id=CM000679.2>
##region-info=<name=MAPT,id=GL000258.2,assoc_id=CM000679.2,reg=45309498-46836265>
##INFO=<ID=ALTLOC,Number=.,Type=String,Description="A list of the alternate loci is the reference genome that are associated with this locus">
##INFO=<ID=ALTHAP,Number=.,Type=String,Description="A list of the known haplotypes that are associated with this locus">

#CHROM      POS REF ALT   INFO                                    FORMAT    NA10847   NA12878   NA21599
chr17       111 A   T     ALTLOCS=GL000258.2;ALTHAPs=H1,H2,H2D    GT:HT     0/1:0/2   0/0:0/0   1/1:2/2
chr17       222 G   A     ALTLOCS=GL000258.2;ALTHAPs=H1,H2,H2D    GT:HT     0/1:0/2   0/0:0/0   1/1:2/2
chr17       333 A   G     ALTLOCS=GL000258.2;ALTHAPS=H1,H2,H2D    GT:HT     0/1:0/2   0/0:0/0   1/1:2/2
. . .
GL000258.2  111 A   T     ALTLOCS=chr17;ALTHAPs=H1,H2,H2D         GT:HT     0/1:0/2   0/0:0/0   1/1:2/2
GL000258.2  222 G   A     ALTLOCS=chr17;ALTHAPs=H1,H2,H2D         GT:HT     0/1:0/2   0/0:0/0   1/1:2/2
GL000258.2  333 A   G     ALTLOCS=chr17;ALTHAPS=H1,H2,H2D         GT:HT     0/1:0/2   0/0:0/0   1/1:2/2
. . .
```

sample	chrom	start	end	hap1	hap2	markers
NA10847	chr17	45309498	46836265	**H1**	**H2D**	rs241039,rs434428,…
NA12878	chr17	45309498	46836265	**H1**	**H1**	rs241039,rs434428,…
NA21599	chr17	45309498	46836265	**H2D**	**H2D**	rs241039,rs434428,…

The good and the bad

Good

- No burden on existing variant callers to adapt to calling w.r.t. alt loci
- Tool for updating VCF can be updated and improved in parallel with variant callers.

Bad

- One more step / file in the variant interpretation pipeline
- This strategy is only applicable to cases where informative markers exist. Use CNVs in WGS: e.g., KANSL1 partial duplications to distinguish MAPT alt loci

The good and the bad

Good

- No burden on existing variant callers to adapt to calling w.r.t. alt loci
- Tool for updating VCF can be updated and improved in parallel with variant callers.

Bad

- One more step / file in the variant interpretation pipeline
- This strategy is only applicable to cases where informative markers exist. Use CNVs in WGS: e.g., KANSL1 partial duplications to distinguish MAPT alt loci

To Do / Discuss

- How best to improve alignment strategies to facilitate variant and haplotype detection?
- How to best represent the resolved alternate loci in VCF format?

Many thanks for helpful discussions with:

Deanna Church
Brad Holmes
Karyn Meltz-Steinberg
Heng Li
Colin Hercus