HLA typing with build 38 and OptiType

Brad Chapman, Miika Ahdesmaki, Justin Johnson
AstraZeneca Translational Oncology
Bioinformatics Core, Harvard Chan School

10 February 2016

HLA typing

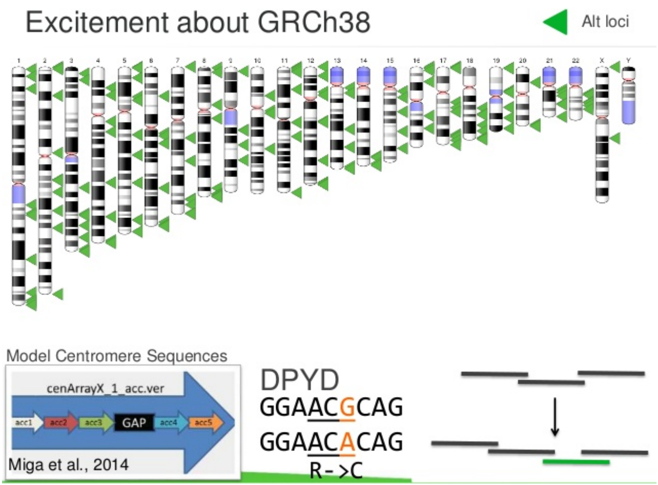
- 1000 genomes: build 38 + IMGT/HLA-3.18.0
- bwa mem extracts HLA reads
- Map reads only to HLA sequences
- OptiType: Call HLA types

https://github.com/lh3/bwa/blob/master/README-alt.md#hla-typing

https://github.com/FRED-2/OptiType

https://github.com/chapmanb/bcbio-nextgen

GRCh38 – graph based, many more alternative loci



Excitement about GRCh38

Model Centromere Sequences

Miga et al., 2014

DPYD

GGAACGCAG

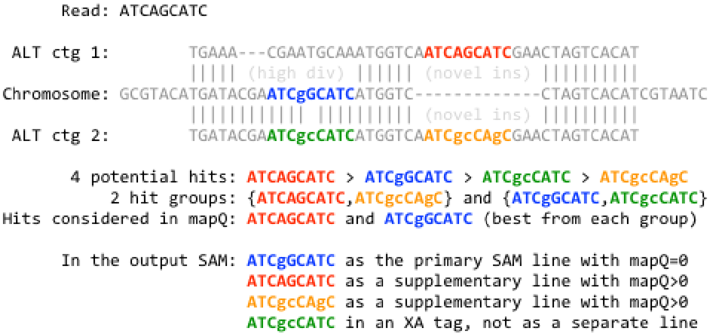
GGAACACAG

R->C

http://www.slideshare.net/GenomeRef/transitioning-to-grch38

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/

Alignment: bwa alternative allele support



```
      Read: ATCAGCATC

 ALT ctg 1:        TGAAA---CGAATGCAAATGGTCAATCAGCATCGAACTAGTCACAT
                   |||||  (high div)  |||||| (novel ins) ||||||||||
Chromosome: GCGTACATGATACGAATCgGCATCATGGTC-------------CTAGTCACATCGTAATC
                   |||||||||||| |||||||||| (novel ins) ||||||||||
 ALT ctg 2:        TGATACGAATCgcCATCATGGTCAATCgcCAgCGAACTAGTCACAT

        4 potential hits: ATCAGCATC > ATCgGCATC > ATCgcCATC > ATCgcCAgC
          2 hit groups: {ATCAGCATC,ATCgcCAgC} and {ATCgGCATC,ATCgcCATC}
Hits considered in mapQ: ATCAGCATC and ATCgGCATC (best from each group)

      In the output SAM: ATCgGCATC as the primary SAM line with mapQ=0
                         ATCAGCATC as a supplementary line with mapQ>0
                         ATCgcCAgC as a supplementary line with mapQ>0
                         ATCgcCATC in an XA tag, not as a separate line
```

https://github.com/lh3/bwa/blob/master/README-alt.md

OptiType

- Map reads to HLA exome 2 and 3 from IMGT
- Matrix of sequence matches to alleles
- Formulate as integer linear program (ILP)
- Use ILP solver, like GNU Linear Programming Kit (GLPK)

https://github.com/FRED-2/OptiType

http://bioinformatics.oxfordjournals.org/content/30/23/3310

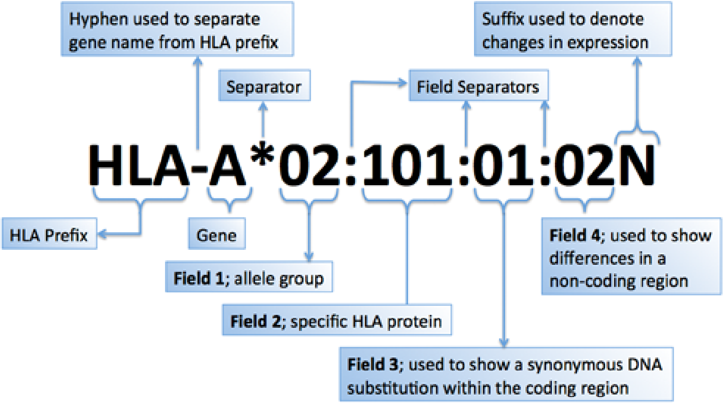
Validations

- Omixon example data
- Exome (1000 genomes) and deep targeted data
- HLA type I calls (A, B, C)
- Good validation results
  - 24/24 (100%) on targeted
  - 22/24 (92%) on exome

http://www.omixon.com/hla-typing-example-data/

https://gist.github.com/chapmanb/8f994618a7fc5e88f893

HLA P-group resolution


HLA-A*02:101:01:02N
Hyphen used to separate gene name from HLA prefix
Separator
Field Separators
Suffix used to denote changes in expression
HLA Prefix
Gene
Field 1; allele group
Field 2; specific HLA protein
Field 3; used to show a synonymous DNA substitution within the coding region
Field 4; used to show differences in a non-coding region

© SGE Marsh 04/10

https://www.ebi.ac.uk/ipd/imgt/hla/

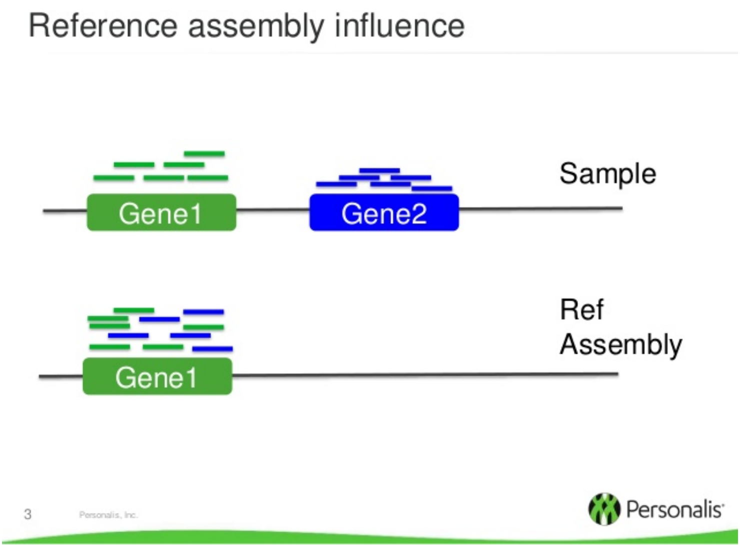
http://hla.alleles.org/alleles/p_groups.html

build 38 validation

- Build 37 and 38
- Validation sets: Genome in a Bottle, Illumina Platinum Genomes
- Lift-over methods: CrossMap/LiftOver, NCBI Remap
- 38 builds: with/without alternative alleles
- Variant callers: FreeBayes, GATK HaplotypeCaller

http://bcb.io/2015/09/17/hg38-validation/

GRCh38 – advantage for variant calling



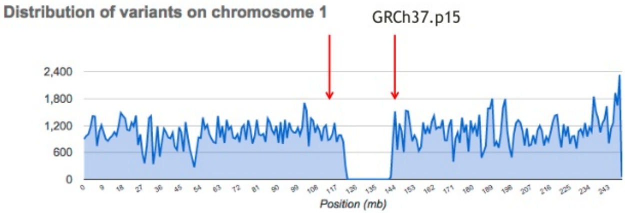
Reference assembly influence

3 Personalis, Inc.

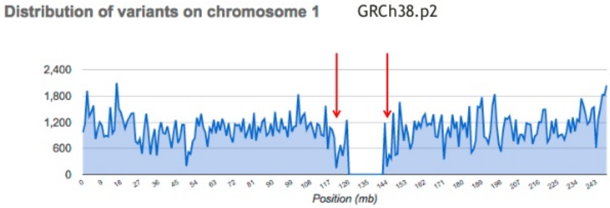
Personalis

http://www.slideshare.net/GenomeRef/transitioning-to-grch38

Avoiding collapsed repeats



GRCh37.p15



GRCh38.p2

http://www.slideshare.net/kmsteinberg/the-importance-of-high-quality-reference-genome-assemblies-to-personal-and-medical-genomics

Reference materials

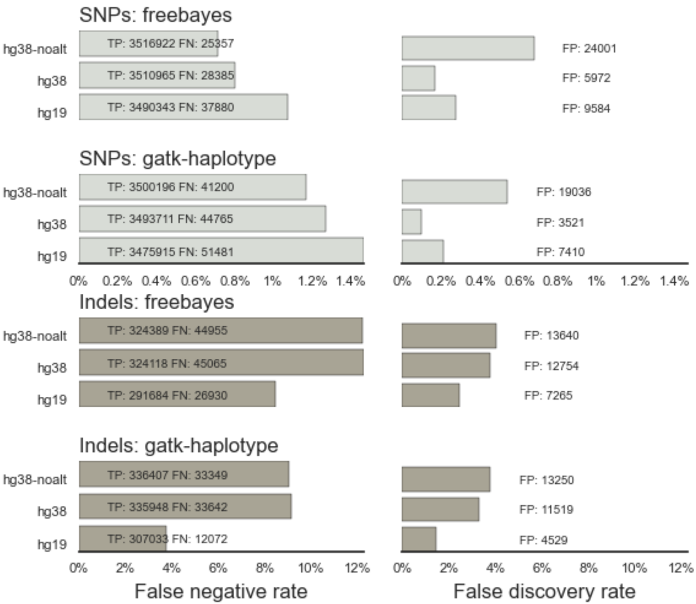
Genome in a Bottle Consortium

Global Alliance for Genomics & Health

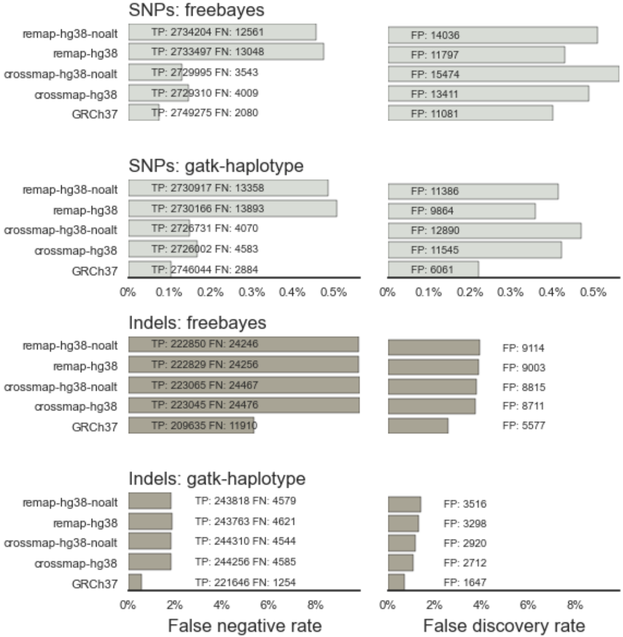
ICGC-TCGA DREAM Mutation Calling challenge

http://www.genomeinabottle.org/
http://ga4gh.org/#/benchmarking-team
https://www.synapse.org/#!Synapse:syn312572

hg19/hg38 comparison: NA12878 Platinum Genomes



GRCh37/hg38 comparison: NA12878 Genome in a Bottle



Small variant results

- SNPs: build 38 more sensitive
- SNPs: build 38 reduces false positives
- Indels: build 38 detected more
- Indels: work on sensitivity and precision

Remapping results

Need conversion approaches for resources not yet available on build 38

- CrossMap: http://crossmap.sourceforge.net/
- NCBI remap: http://www.ncbi.nlm.nih.gov/genome/tools/remap
- Both performed well
- NCBI remap has additional sensitivity, but requires tuning