Analysis of High-Resolution HLA Typing by NGS: Which tools do we need?

Leipzig, 2019

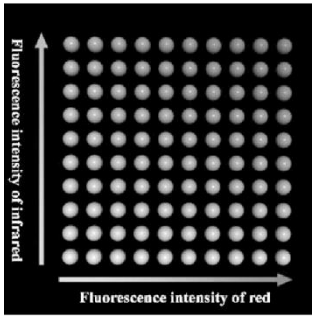
Nina Lauterbach, Senior Field Application Scientist

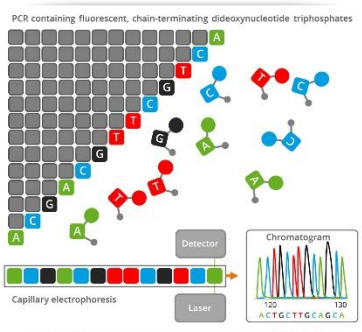
Agenda

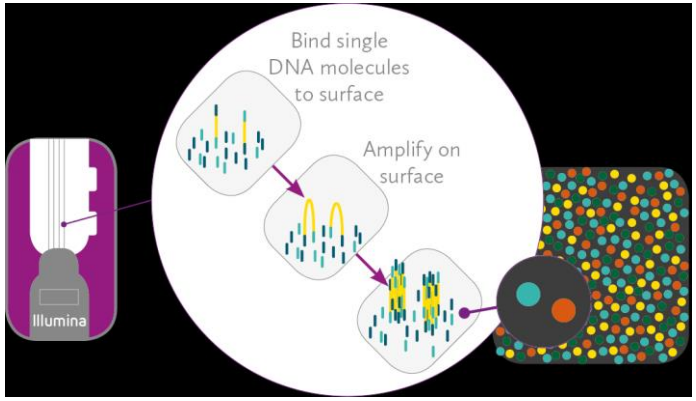
- Brief History
- SBT vs NGS
- Benefits and challenges of NGS
- Tools in the NGS analysis software – HLA Twin
- Summary
- Product update – Omnitype
- Questions

Evolution of molecular HLA typing methods

December 16th 1998:
Introduction of HLA
IMGT database







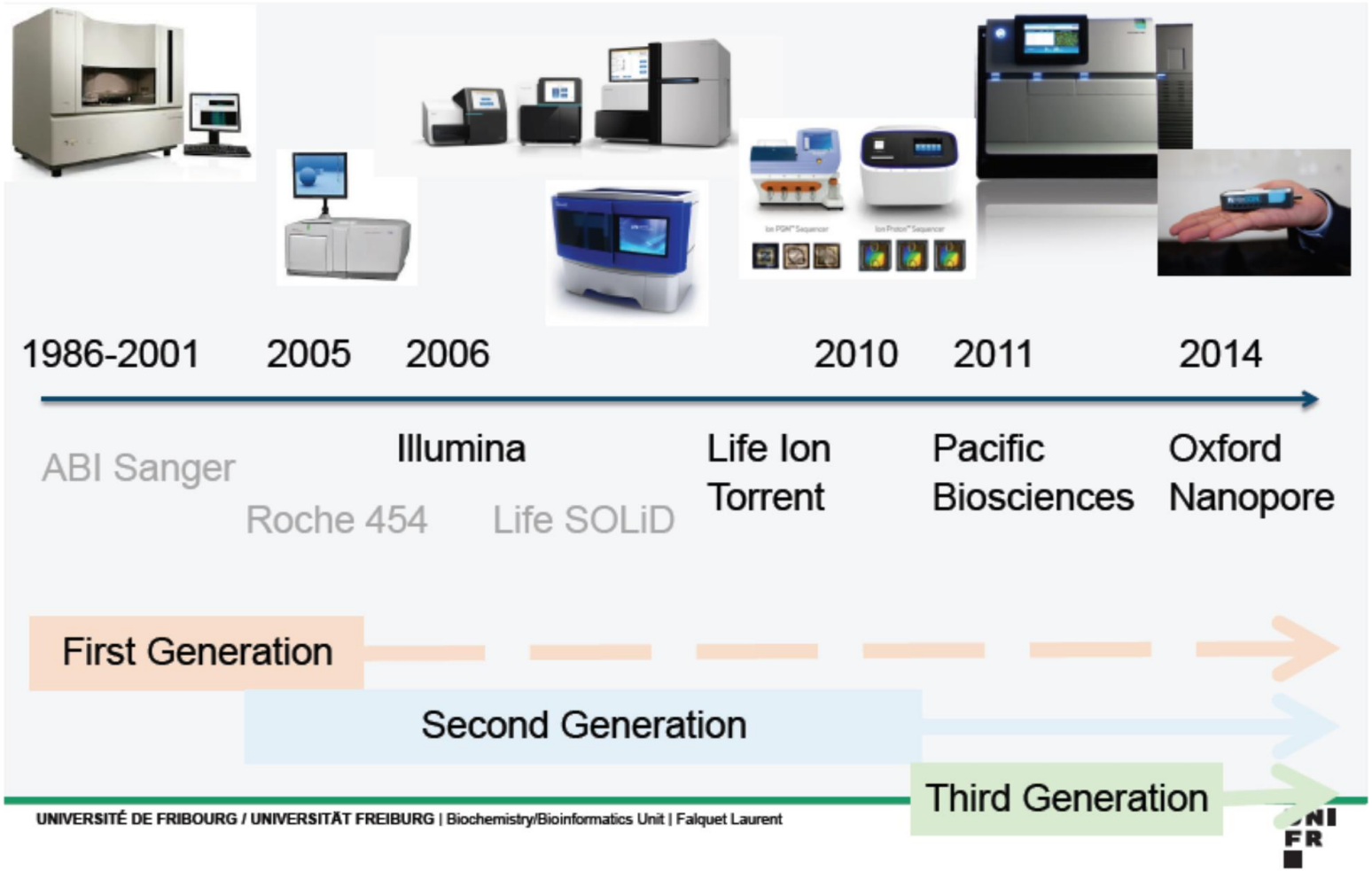
NGS - Illumina

Sanger sequencing

PCR-SSP/SSO

1991 1996 2001 2006 2013 2018

Evolution of NGS platforms



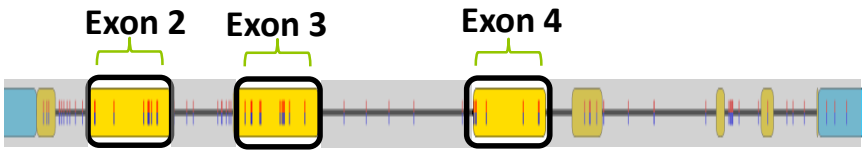
What do we gain from NGS?

1. Full length gene sequencing
2. The ability to identify novel/ null alleles
3. Typing with (almost) zero ambiguities
4. Flexibility in sample number, from low- to high throughput
5. Possibility to automate pre- and post-PCR
6. Cost efficient – possibility to combine different type of libraries

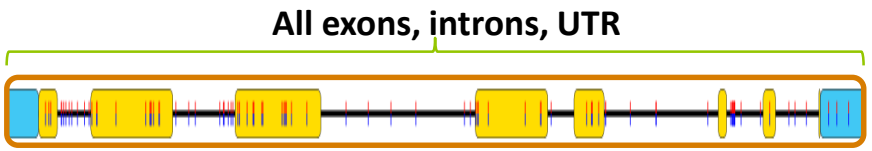
SBT

NGS

HLA-A,B,C genes sequenced at Exons 2,3,4 only (822 bases)

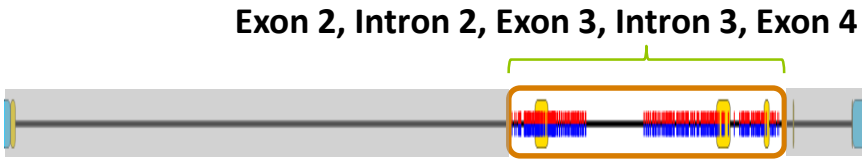


HLA-A,B,C genes sequenced full length (~3,200 bases)

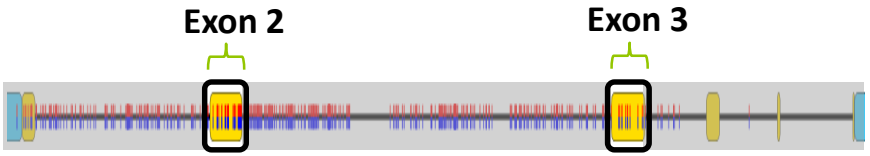


HLA-DRB1 genes sequenced at Exon 2 only (276 bases)

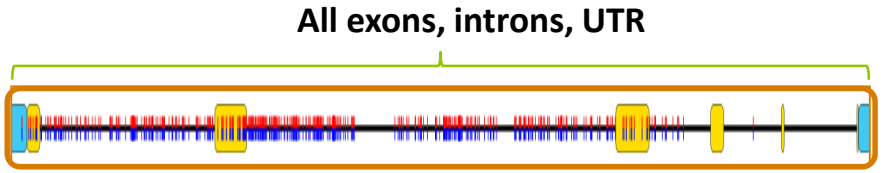
HLA-DRB1 genes sequenced partial length (~5,500 bases)



HLA-DQB1 genes sequenced at Exons 2 and 3 only (540 bases)



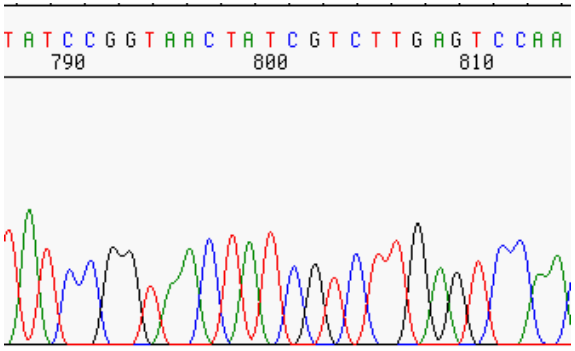
HLA-DQB1 genes sequenced full length (~7,000 bases)

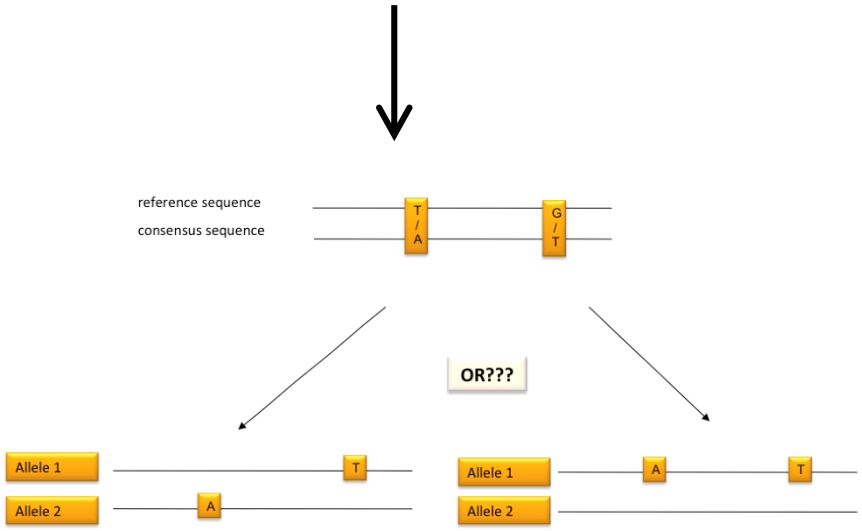


Courtesy of Curt Lind at CHOP

Benefit of phasing with NGS

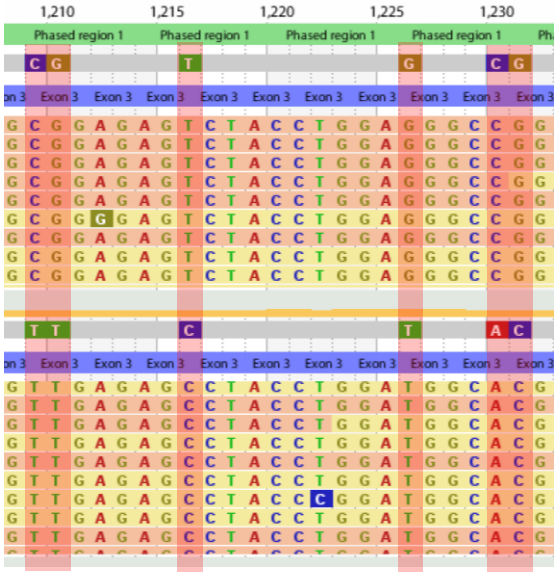
SBT – 1 signal





NGS – 2 signals

Allele	A	C	T	G
HLA-C*P1:C1		135		1
HLA-C*P1:C2		1		193

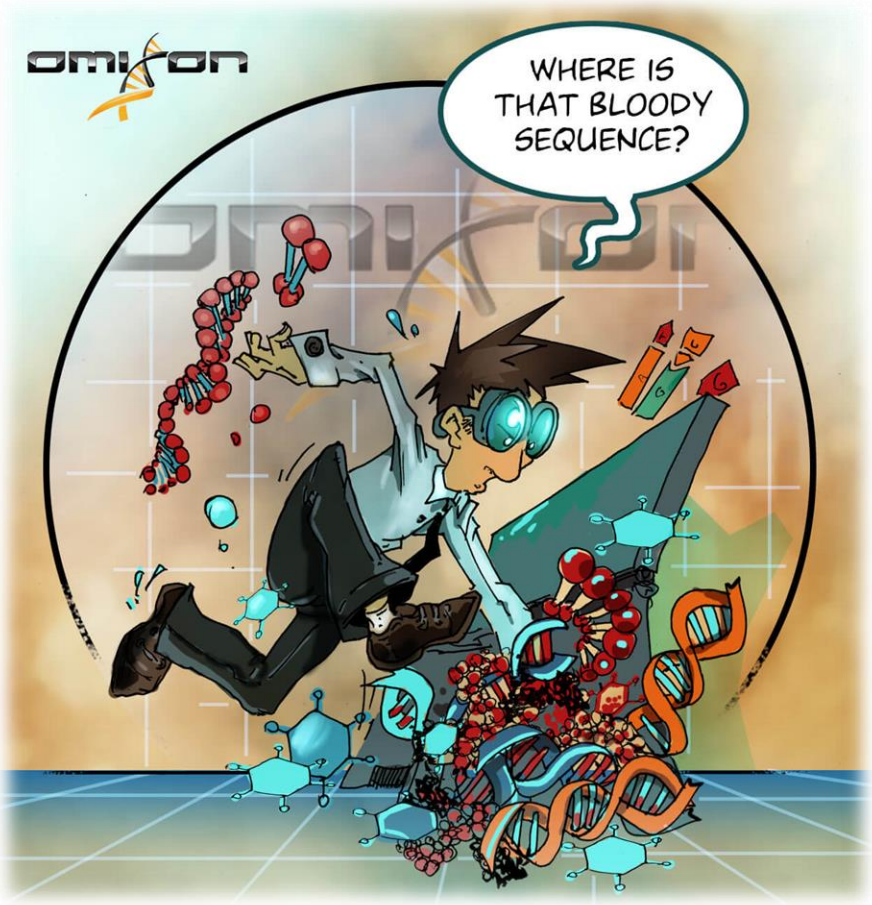


Challenge of HLA typing using NGS

- New definitions (phasing, coverage, Q30 score etc.)
- Complex algorithms
- Incomplete reference allele database
- Increasing allele database (35% increase from 3.33 to 3.38)
- Large and more complex dataset

NGS analysis software

The software is crucial in order to make sense of the huge amount of sequencing data



Illumina as primary platform

ABI 3730
Capillary sequencer
Low volume, high labor

Roche 454 GS Junior
Benchtop NGS
Expensive per run

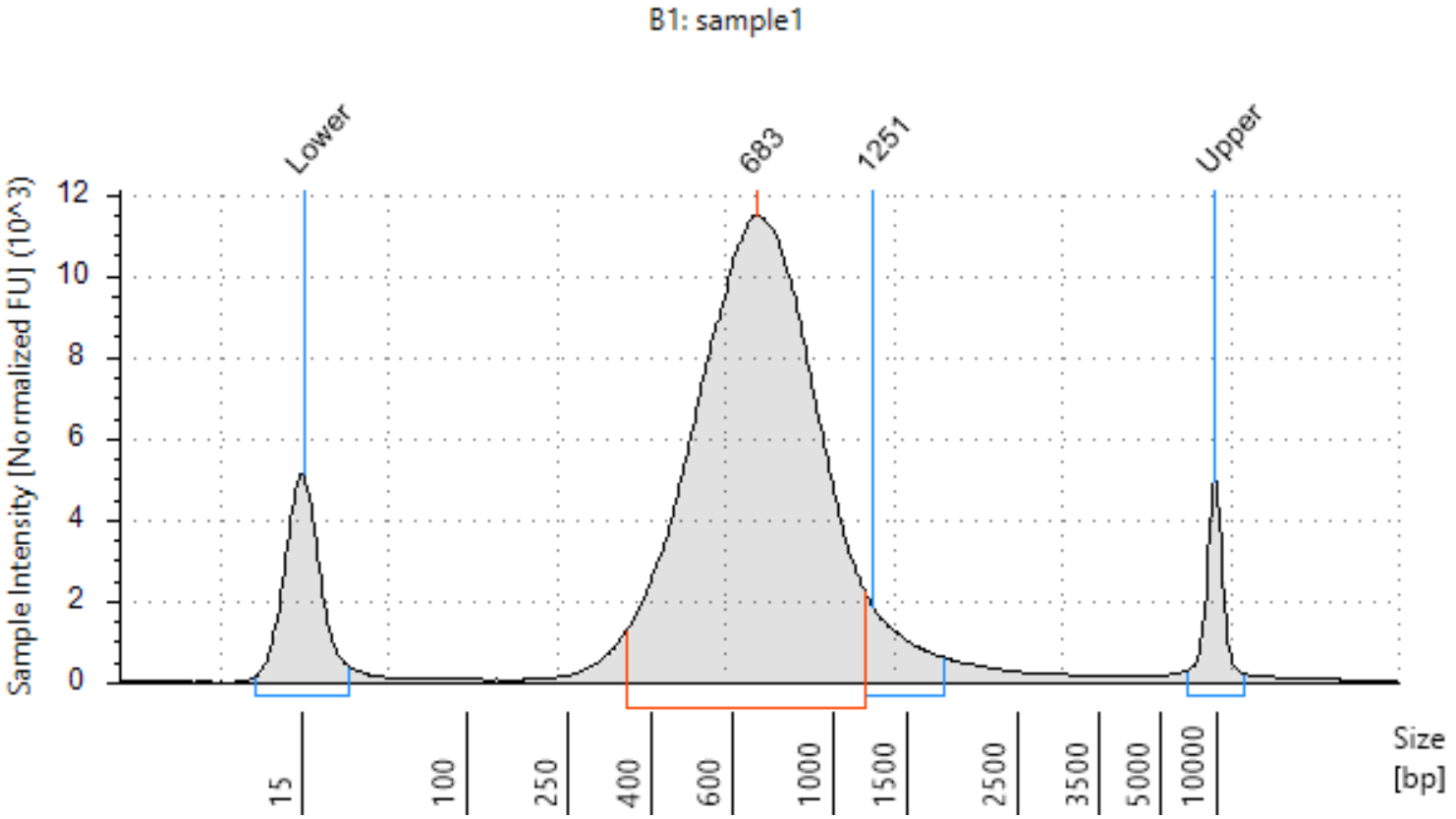
Genestudio S5
Benchtop NGS
Semi-conductor sequencing
Long sample prep
Rapid run time

PacBio RS
Non Benchtop (large)
Fluorophore photocapture
S M R T
High raw error

Illumina MiSeq
Benchtop NGS
Highly automated
Quick sample prep
Lowest error rate

Library contains various fragment sizes

Fragment distribution after the bead-based size selection (500 – 1500 bp)



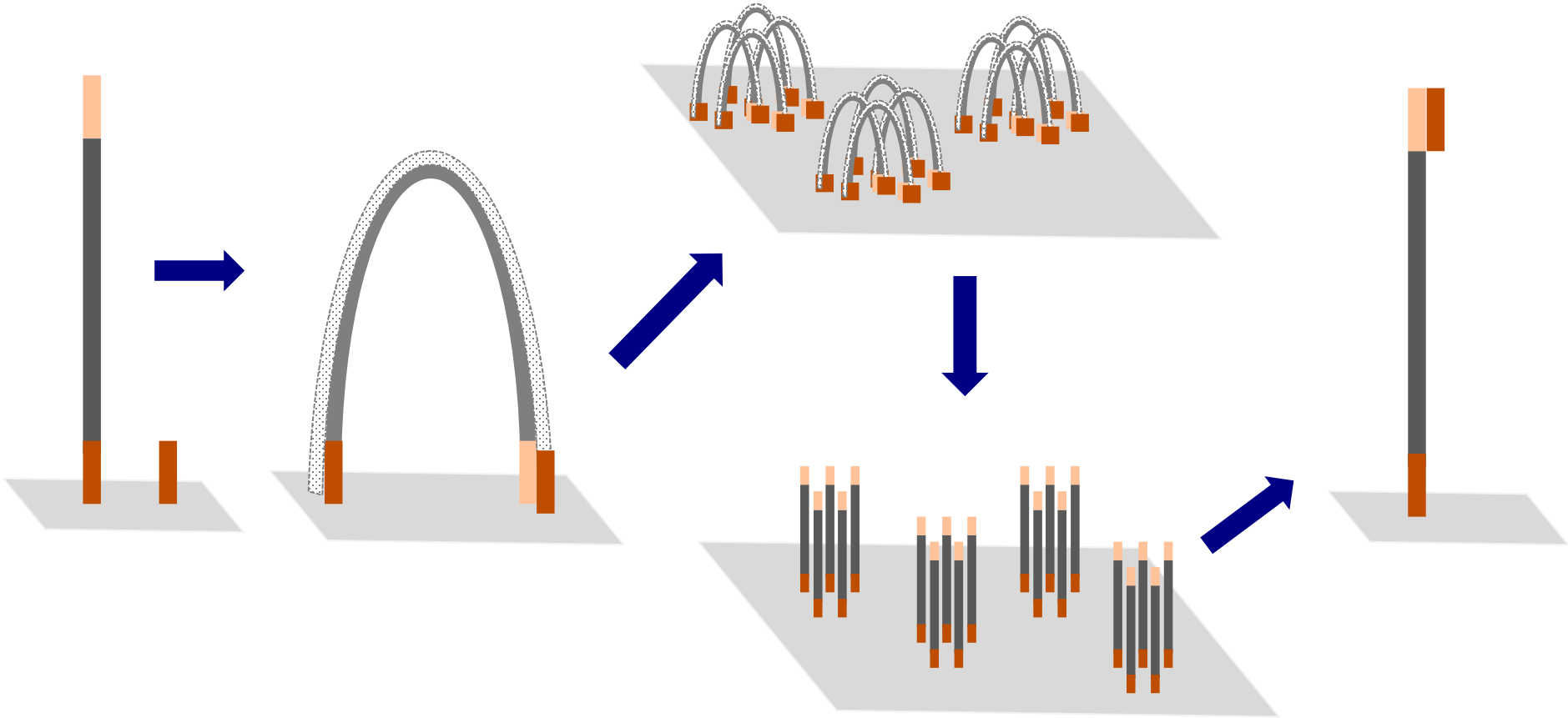
Illumina NGS Chemistry

Attach DNA to flow cell

Perform bridge amplification

Generate clusters

Anneal sequencing primer

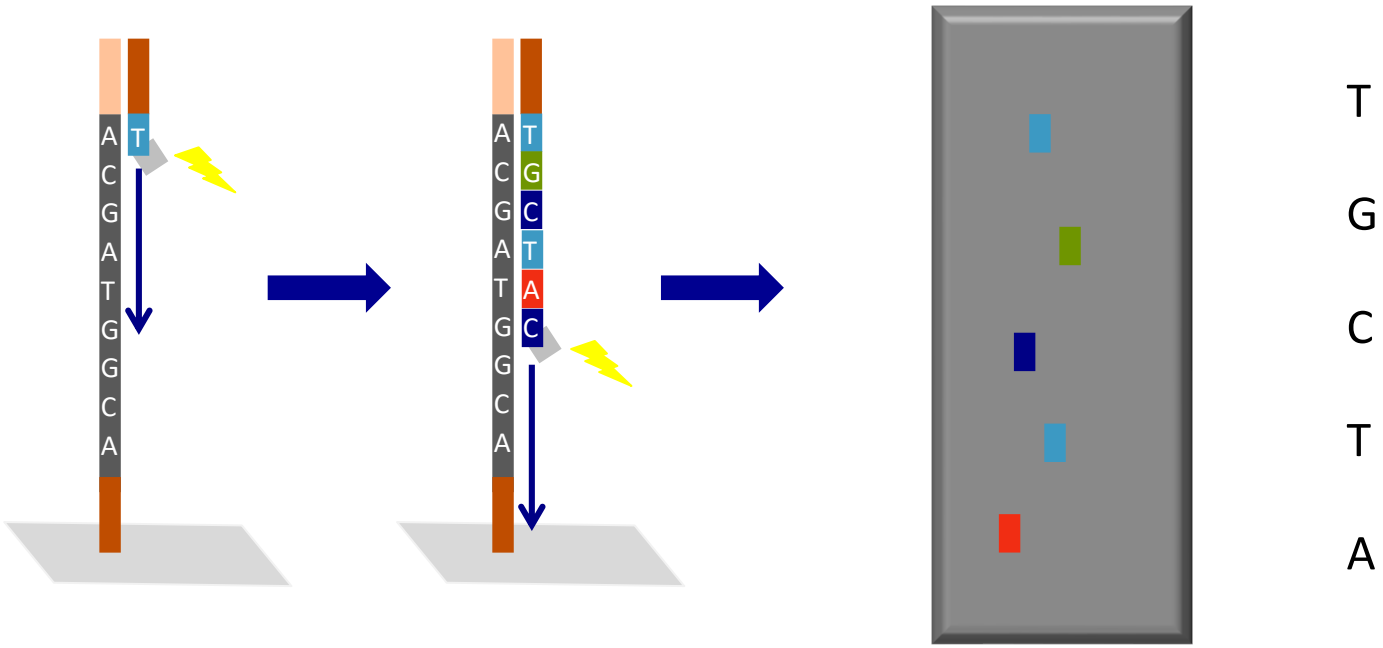


Illumina NGS Chemistry

Extend first base and read it

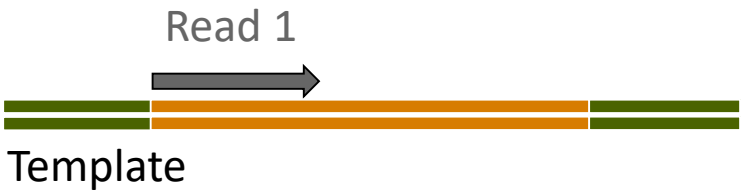
Repeat previous step to extend strand

Generate base calls

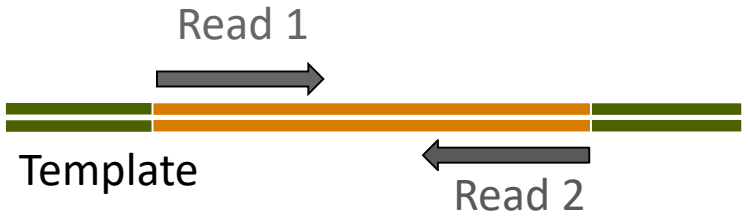


Paired-End Sequencing

Single-End Reads

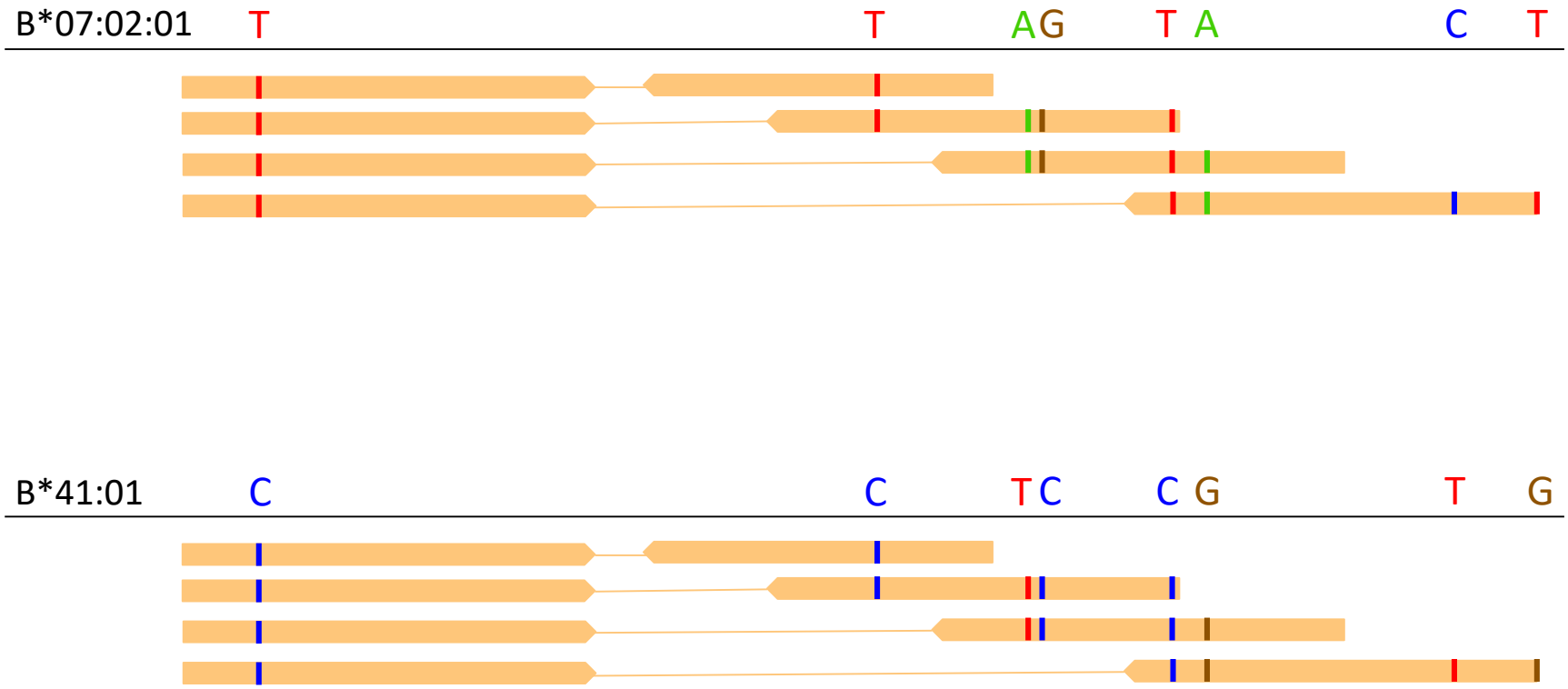


Paired-End Reads

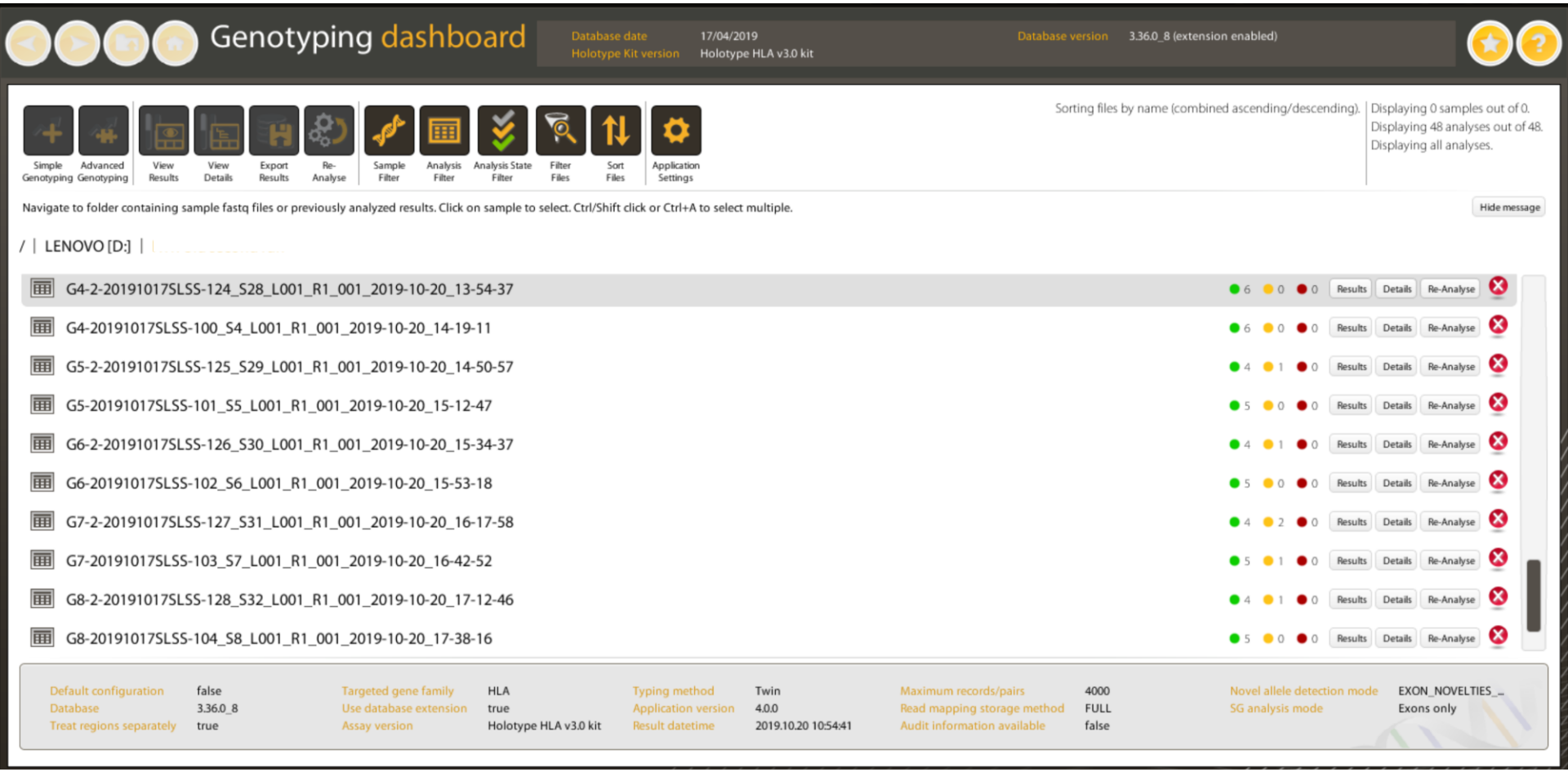


Sequence from the paired reads originate from the same template, and therefore are phased with each other

Long Distance Phasing



Analysis paired-end reads with HLA Twin

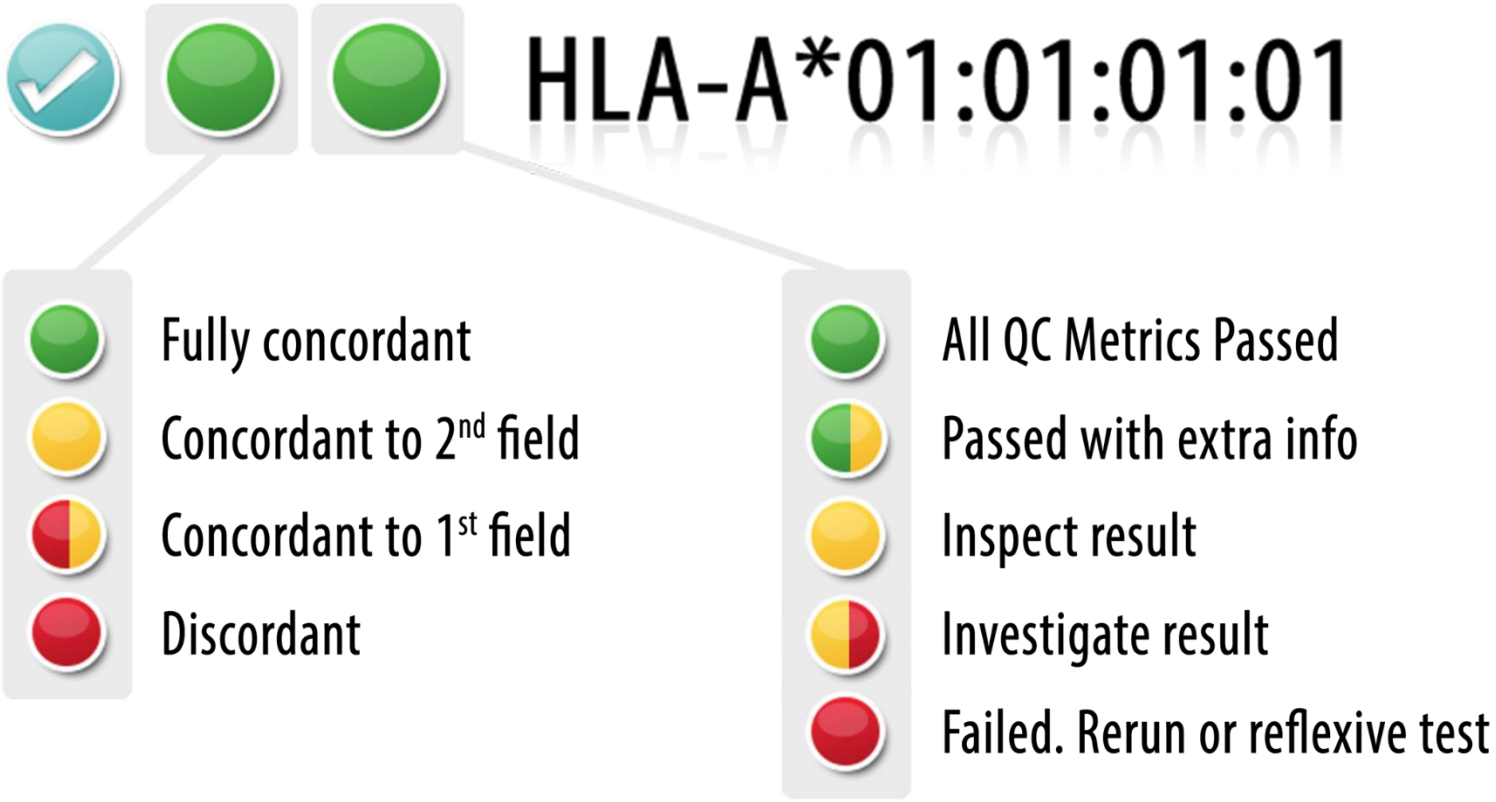


Omixon HLA Twin (Software)

- Preconfigured Holotype HLA-specific settings
- Automated genotyping after MiSeq run
- Up to two algorithms for determining HLA genotypes
  - Consensus Genotyping (Assembly)
  - Statistical Genotyping (Alignment to IMGT/HLA)
- Traffic light system for data interpretation and workflow management

Passed
Info
Inspect
Investigate
Failed

HLA Twin - The Traffic Light System



HLA Twin

- Two algorithms for determining HLA genotypes:
  1. Consensus Genotyping (Assembly)
  2. Statistical Genotyping (Alignment to IMGT/HLA)

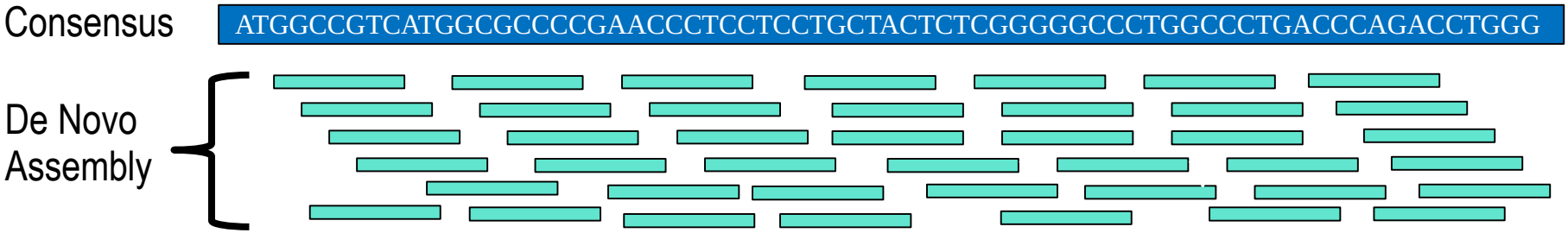
Consensus Genotyping (Assembly) :

1. C- collect
2. C- construct
3. C- compare
4. C- check

HLA Twin - Consensus algorithm (1)

1. C- collect - The HLA Twin will gather all sequencing reads from the Illumina sample file pair and will assign them to a specific locus.

1. C- Construct - The processed reads per loci will be assembled (De Novo Assembly) to construct a long and continuous sequence called the consensus.

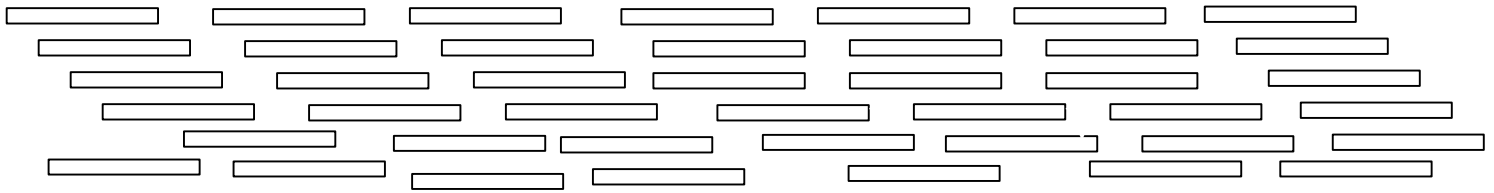


> During the process of De Novo Assembly, a filtering algorithm will detect reads with too low quality, which will be filtered out.

HLA Twin - Consensus algorithm (2)

3. C- Compare After completion, the consensus sequence will be compared to all the reference sequences of the alleles (of that specific loci) in the IMGT/HLA database to find a best match allele and to make an allele call.

Consensus ATGGCCGTCATGGCGCCCCGAACCCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGACCCAGACCTGGG



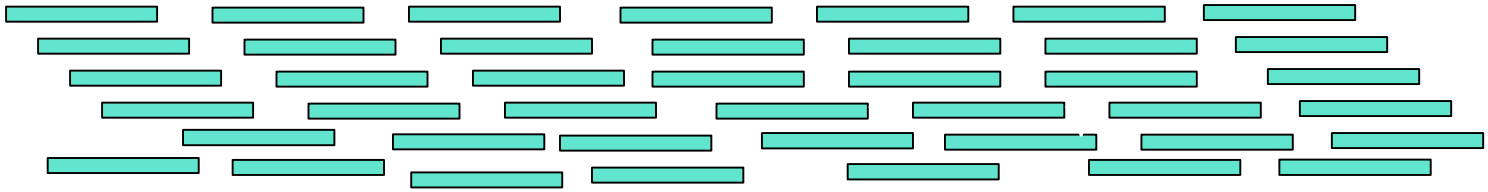
cDNA	10	20	30	40	50	60	70
A*01:01:01:01	ATGGCCGTCA	TGGCGCCCCG	AACCCTCCTC	CTGCTACTCT	CGGGGGCCCT	GGCCCTGACC	CAGACCTGGG
A*01:01:01:02N	----------	----------	----------	----------	----------	----------	----------
A*01:01:01:03	----------	----------	----------	----------	----------	----------	----------
A*01:01:01:04	----------	----------	----------	----------	----------	----------	----------
A*01:01:01:05	----------	----------	----------	----------	----------	----------	----------
A*01:01:01:06	----------	----------	----------	----------	----------	----------	----------
A*01:01:01:07	----------	----------	----------	----------	----------	----------	----------
A*01:01:01:08	----------	----------	----------	----------	----------	----------	----------

Allele call
(For example: HLA-A*01:01:01:01)

HLA Twin - Consensus algorithm (3)

4. C- Check - Upon the allele call the software will re-check itself (since it used only the consensus sequence to identify the best match):
   All the processed sequencing reads will be aligned to the reference sequence of the best matched allele (in this example HLA-A*01:01:01:01).

```
A*01:01:01:01  ATGGCCGTCA TGGCGCCCCG AACCCTCCTC CTGCTACTCT CGGGGGCCCT GGCCCTGACC CAGACCTGGG
```



> Potential mismatches (novel sequences) between the reference sequence and sequencing reads will be detected and highlighted.

HLA Twin Statistical algorithm

Statistical Genotyping : Alignment to IMGT/HLA

Works in a " if " scenery:

1. C- confirm

Statistical genotyping (SG) algorithm

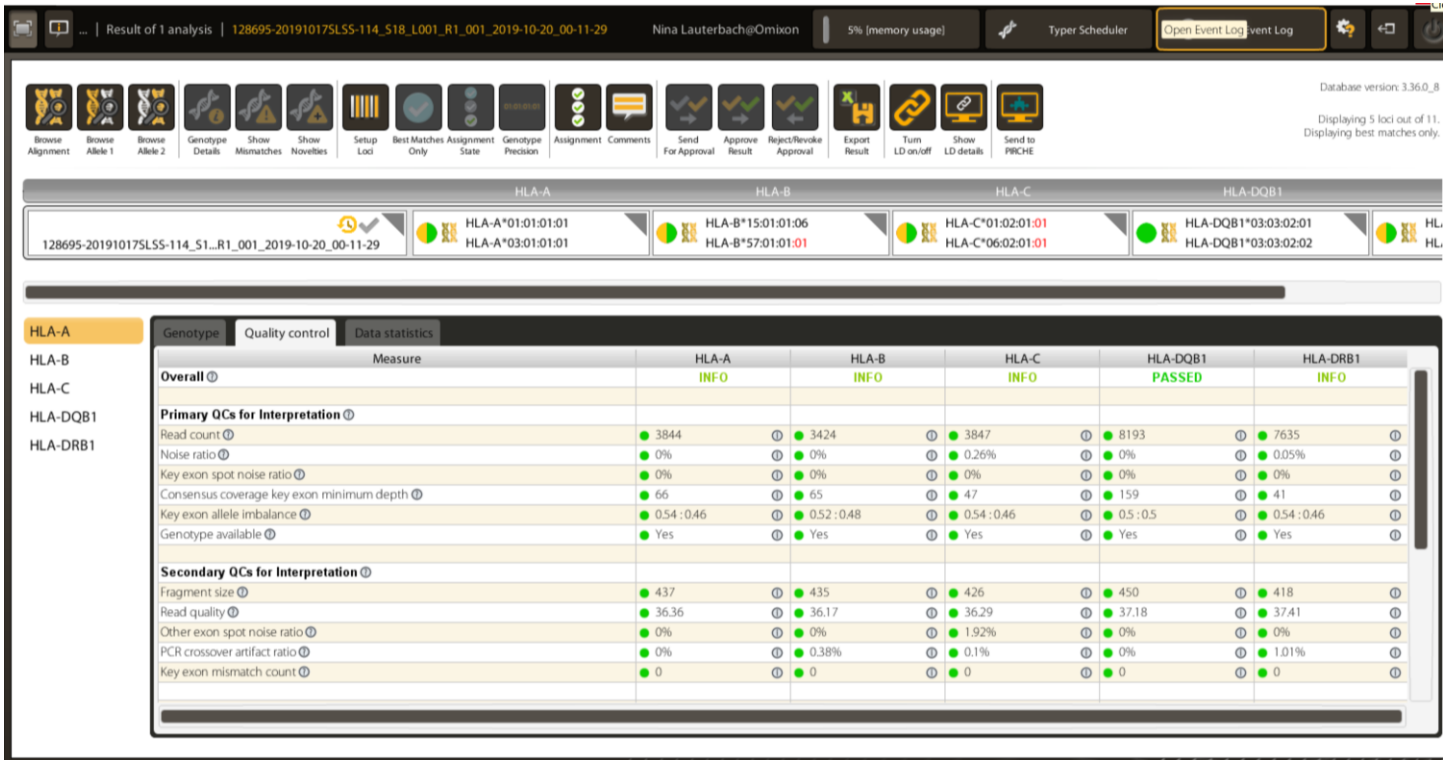
5 – Confirm: If the QC metrics are below a certain threshold (or a novelty is found), the software will try to **confirm the allele call** with the Statistical genotyping.

- The reads from the Illumina sample file pair will be directly **aligned to all the alleles of the IMGT/HLA database**.
- A complex calculation decides which allele has most reads aligned and therefore will be chosen as best match.

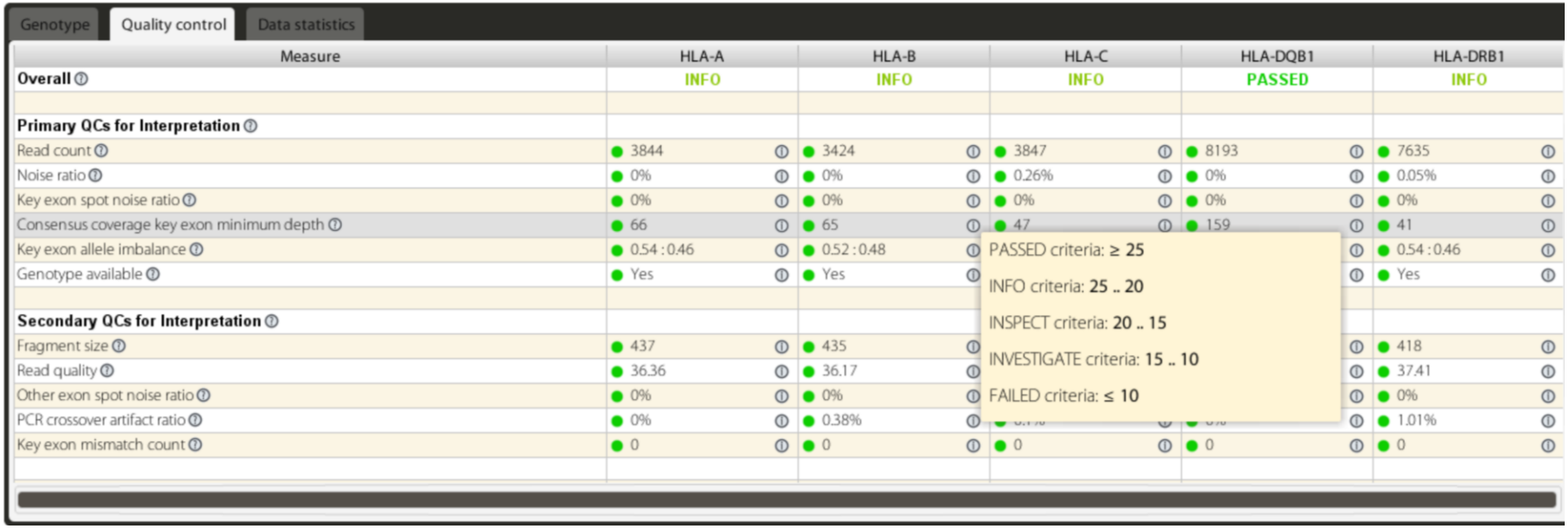
Tools to troubleshoot…

Quality Control Metrics system

- **The 24 quality metrics** will guide you towards the cause of a quality issue and help you decide whether or not you need to repeat the run.



Quality determined according to set thresholds



Genotype | Quality control | Data statistics

Measure	HLA-A	HLA-B	HLA-C	HLA-DQB1	HLA-DRB1
Overall	INFO	INFO	INFO	PASSED	INFO
Primary QCs for Interpretation					
Read count	3844	3424	3847	8193	7635
Noise ratio	0%	0%	0.26%	0%	0.05%
Key exon spot noise ratio	0%	0%	0%	0%	0%
Consensus coverage key exon minimum depth	66	65	47	159	41
Key exon allele imbalance	0.54 : 0.46	0.52 : 0.48			0.54 : 0.46
Genotype available	Yes	Yes			Yes
Secondary QCs for Interpretation					
Fragment size	437	435			418
Read quality	36.36	36.17			37.41
Other exon spot noise ratio	0%	0%			0%
PCR crossover artifact ratio	0%	0.38%			1.01%
Key exon mismatch count	0	0	0	0	0

PASSED criteria: ≥ 25

INFO criteria: 25 .. 20

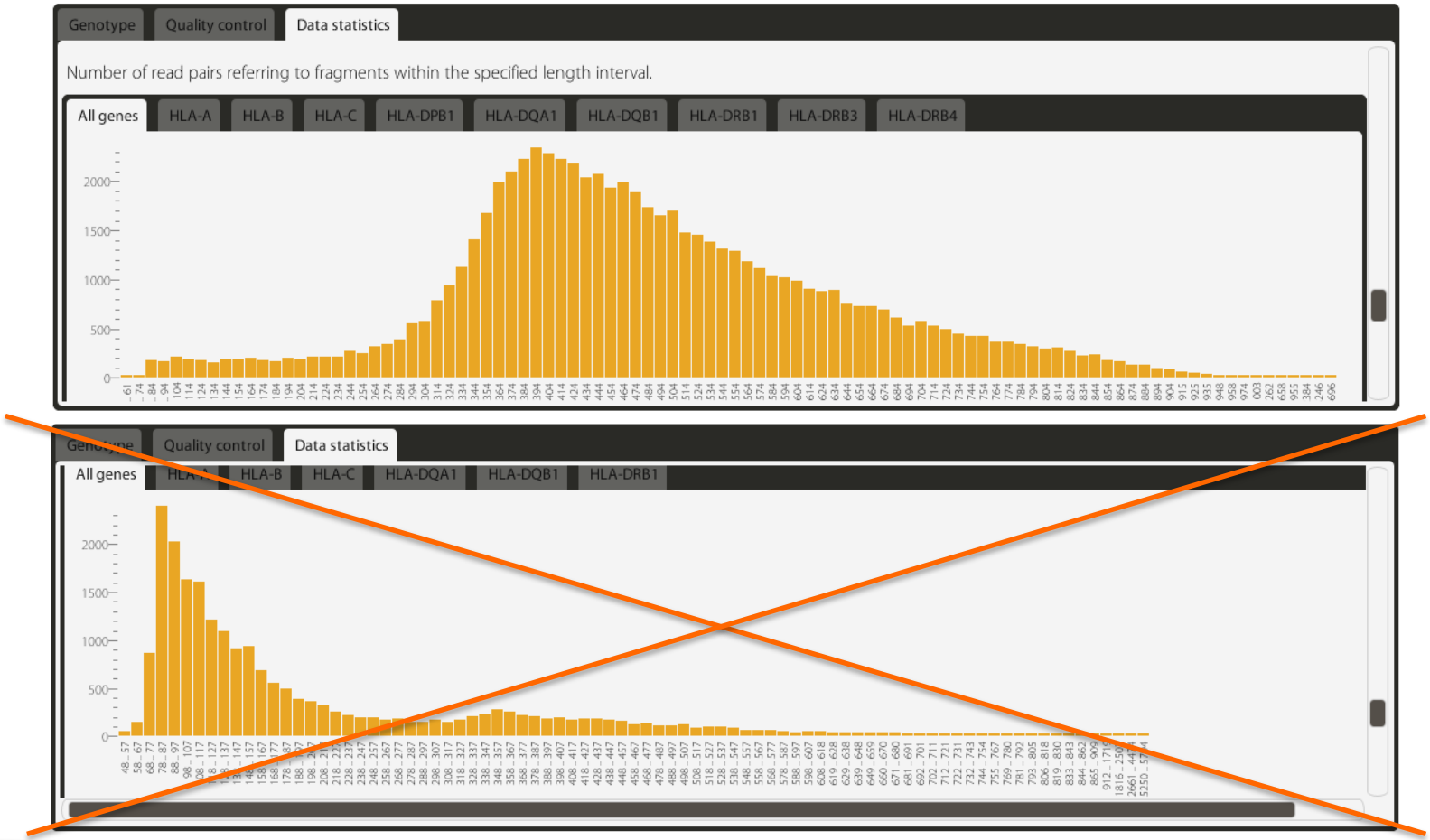
INSPECT criteria: 20 .. 15

INVESTIGATE criteria: 15 .. 10

FAILED criteria: ≤ 10

Fragment size is important

Recommended average fragment size of 400 bp

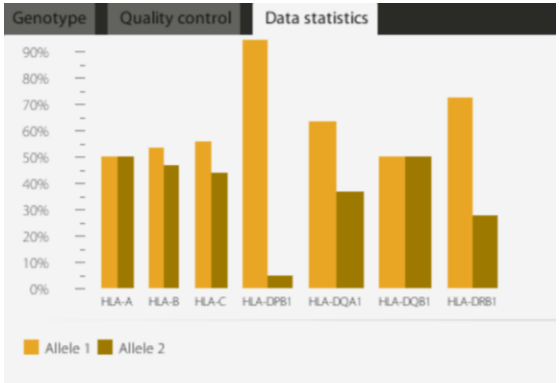


Allele imbalance?

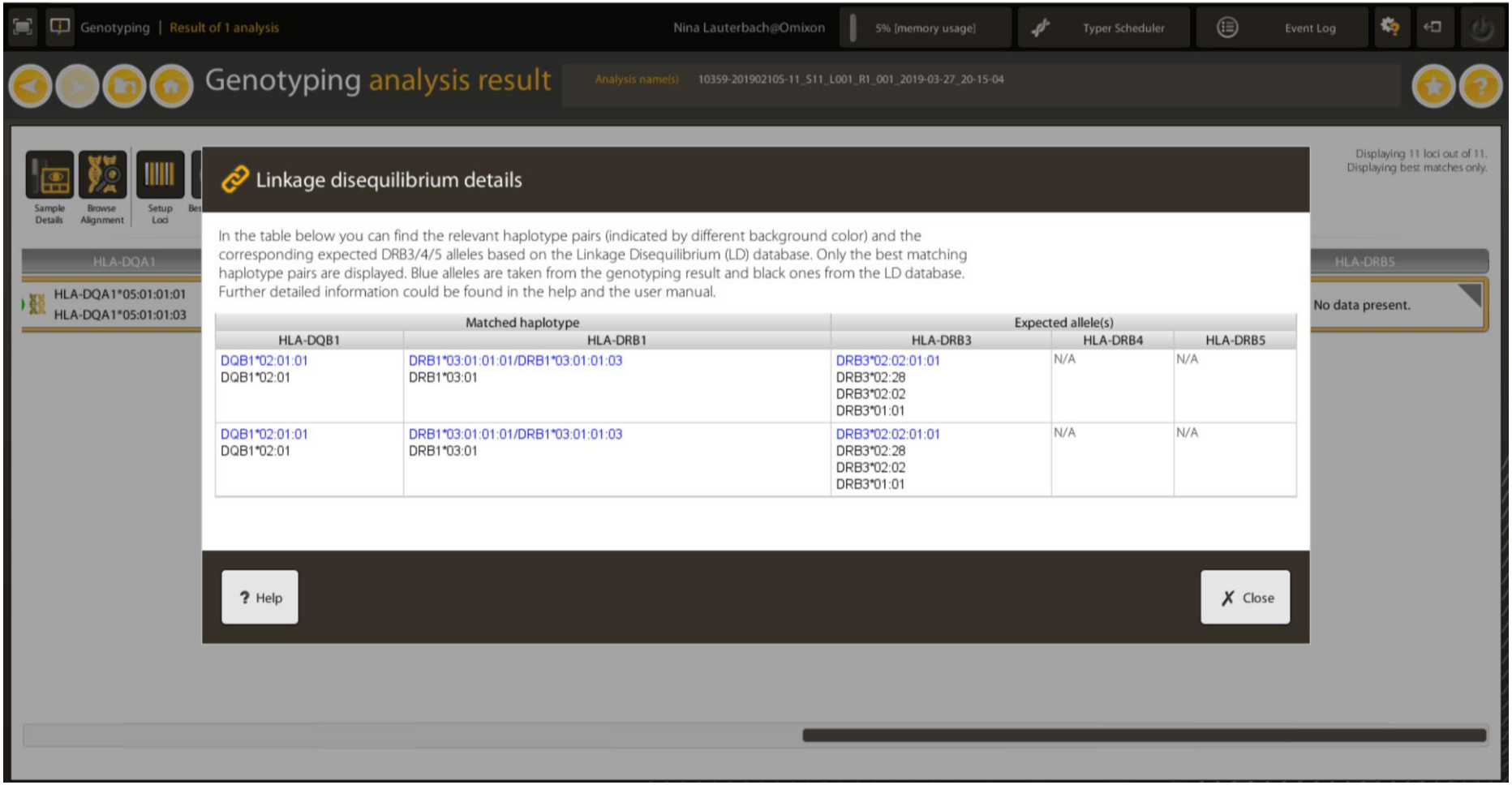
- The software can detect strong allelic imbalance > second allele below 1%

Genotype | Quality control | Data statistics

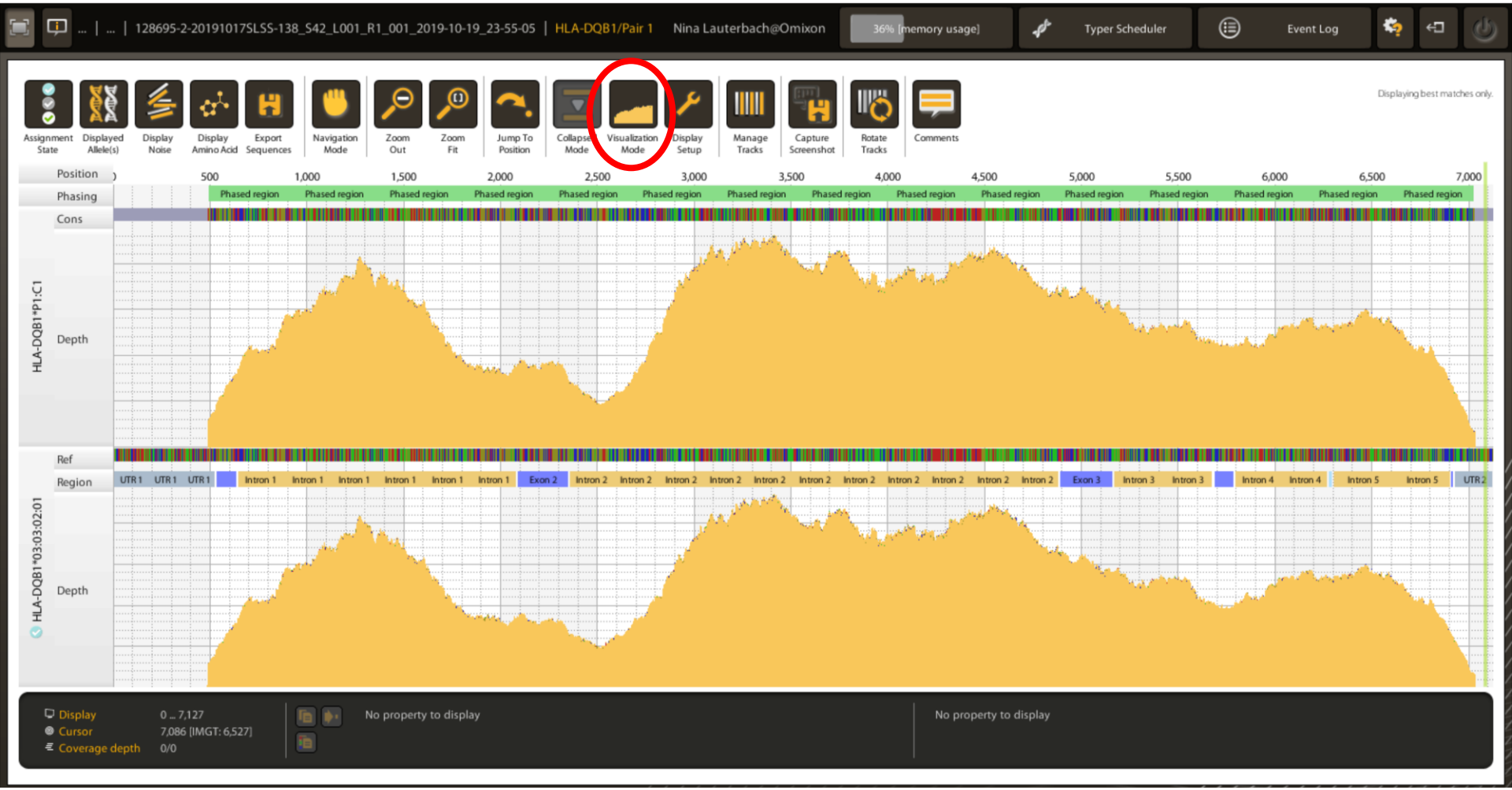
Measure	HLA-A	HLA-B	HLA-C	HLA-DPB1	HLA-DQB1	HLA-DRB1
Key exon mismatch count	0	0	0	0	0	0
Warnings for Troubleshooting						
Read length	144	142	141	145	144	144
Crossmapping (intergenic ambiguity)	10.73%	16.93%	18.8%	0.07%	0.02%	8.39%
Ambiguous layout (intragenic ambiguity)	3.99%	10.71%	13.17%	2.73%	1.35%	1.58%
Non-exon spot noise ratio	0%	7.98%	6.27%	7.45%	4.72%	22.31%
Continuous consensus	Yes	Yes	Yes	Yes	Yes	Yes
Fully phased consensus	Yes	Yes	Yes	No	Yes	Yes
Consensus coverage other exon minimum depth	187	107	83	9	261	197
Consensus coverage non-exon minimum depth	157	64	66	0	30	26
Other exon allele imbalance	0.5 : 0.5	0.55 : 0.45	0.55 : 0.45	0.95 : 0.05	0.5 : 0.5	0.5 : 0.5
Non-exon allele imbalance	0.5 : 0.5	0.53 : 0.47	0.55 : 0.45	0.95 : 0.05	0.5 : 0.5	0.72 : 0.28
Other exon mismatch count	0	0	0	0	0	0
Non-exon mismatch count	0	0	0	0	0	0



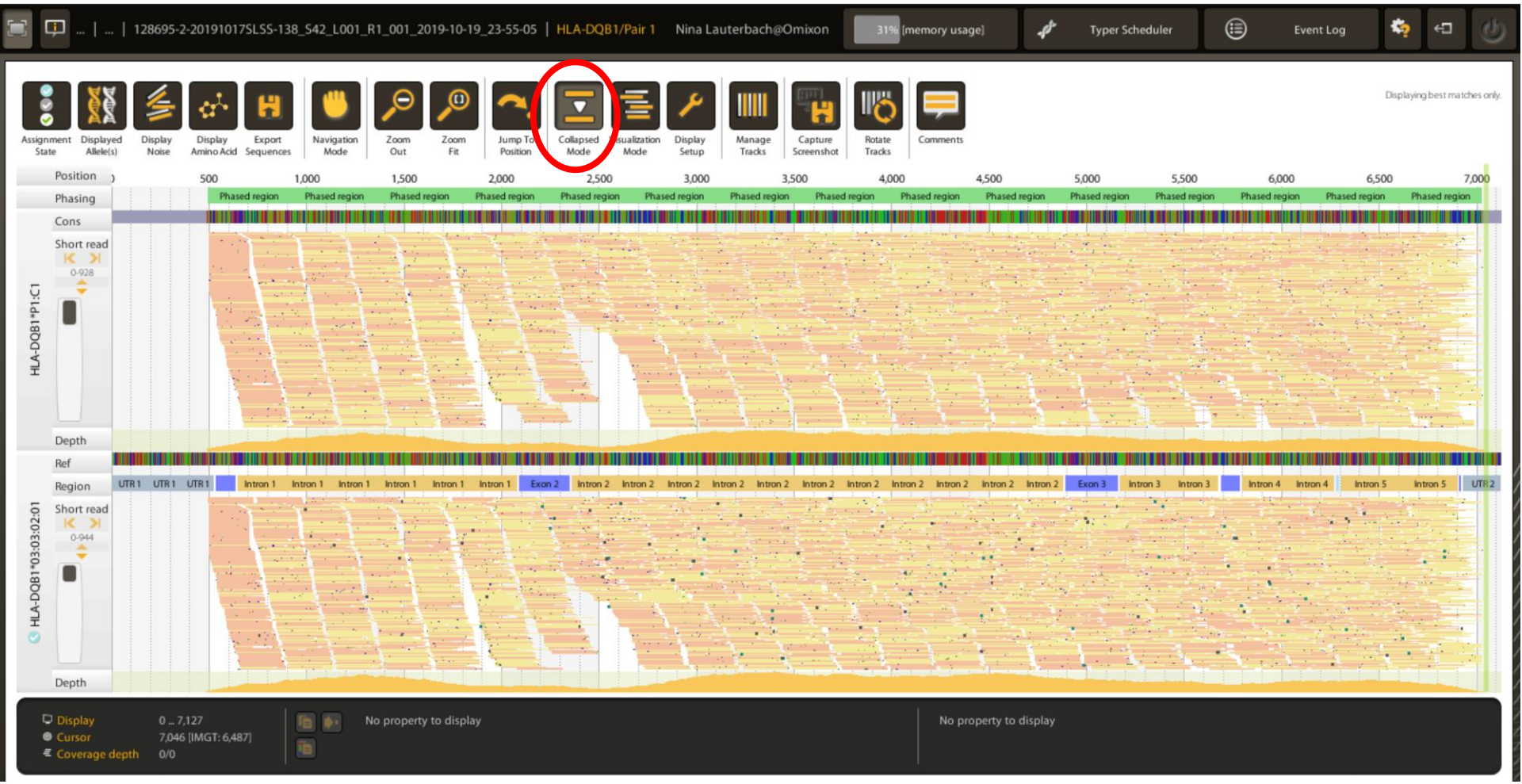
Use of Linkage Disequilibrium database



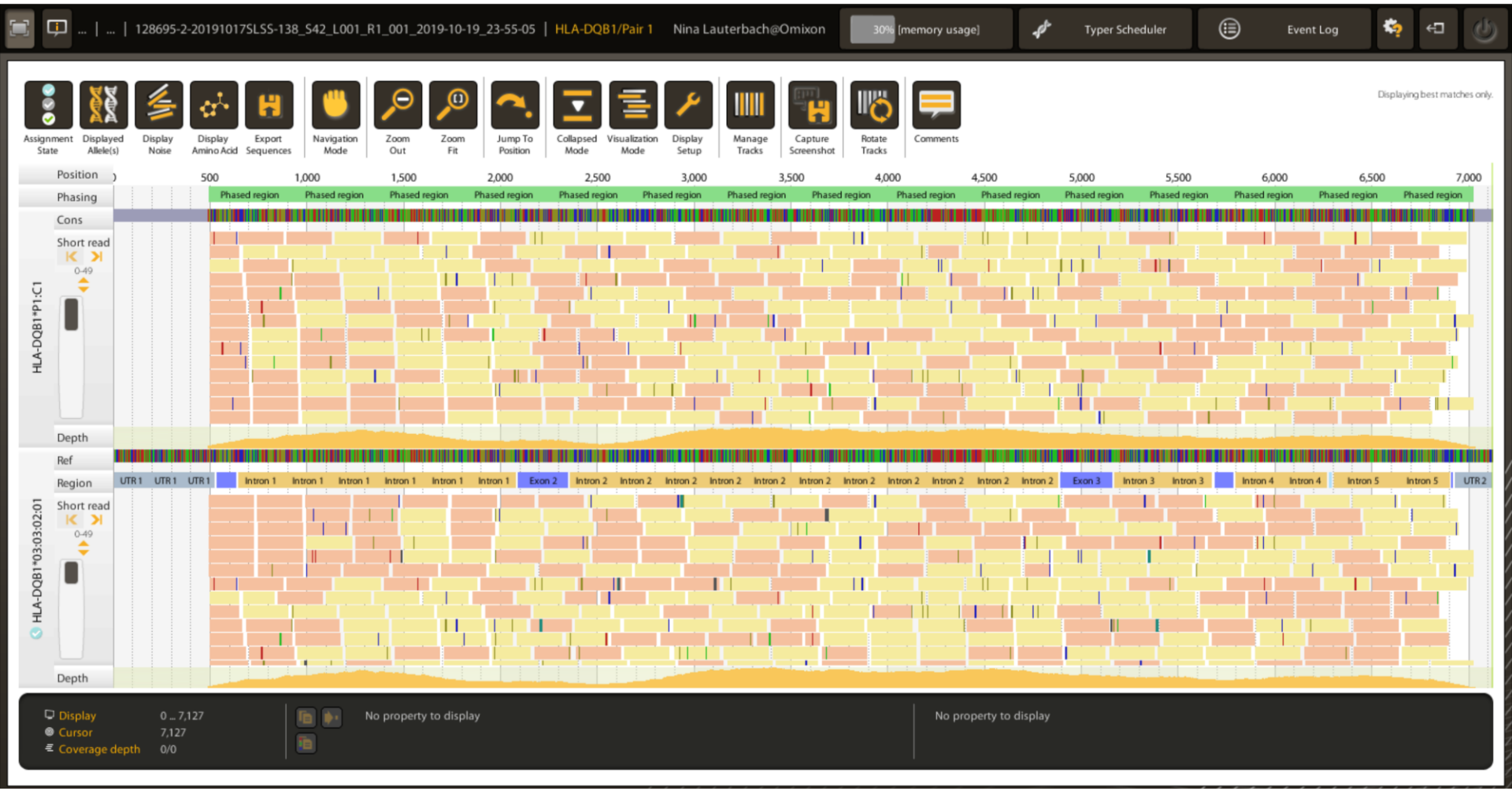
Check the coverage plot in the Gene Browser



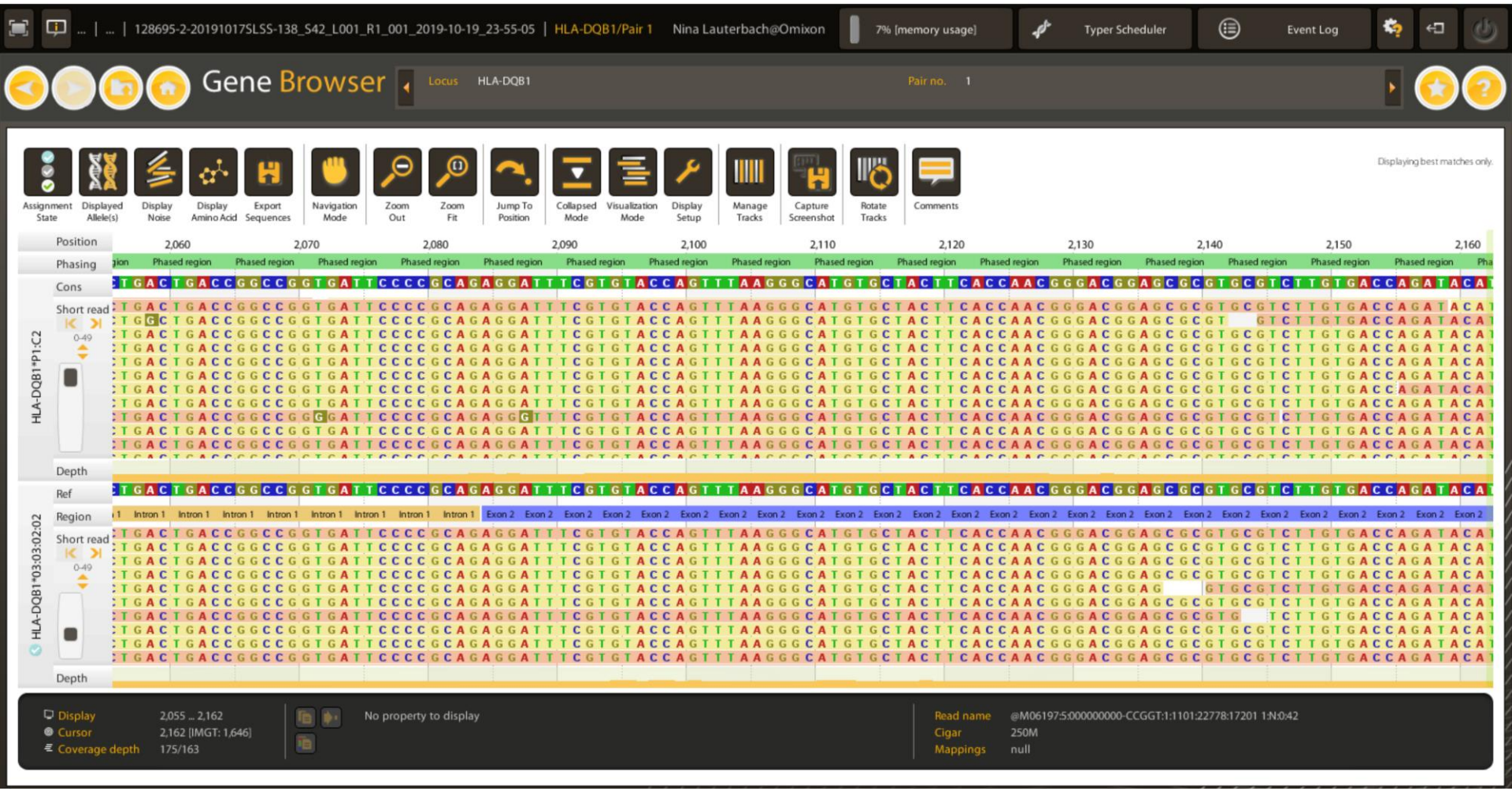
Change the visualization mode



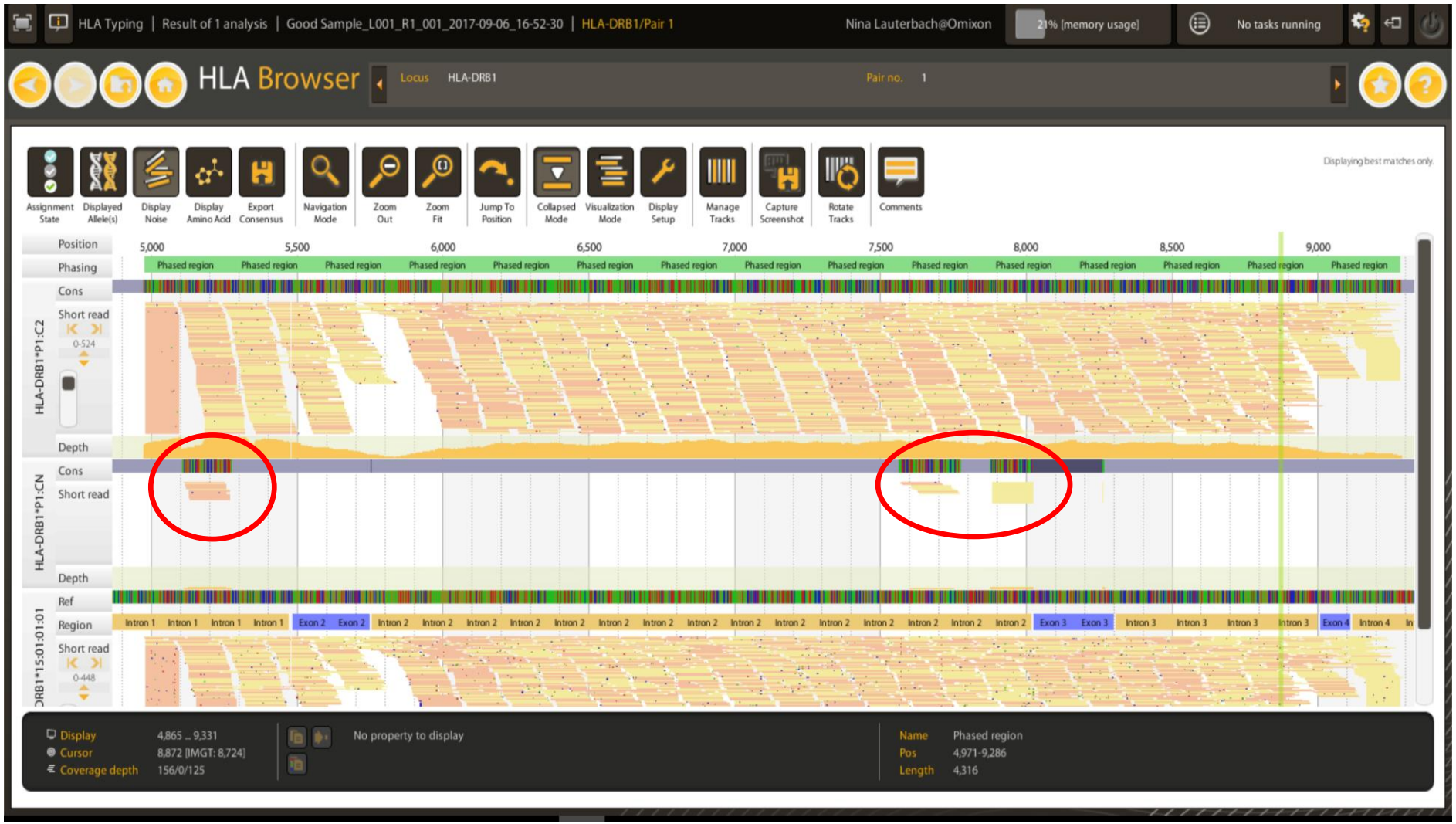
Zoom in on sequencing reads



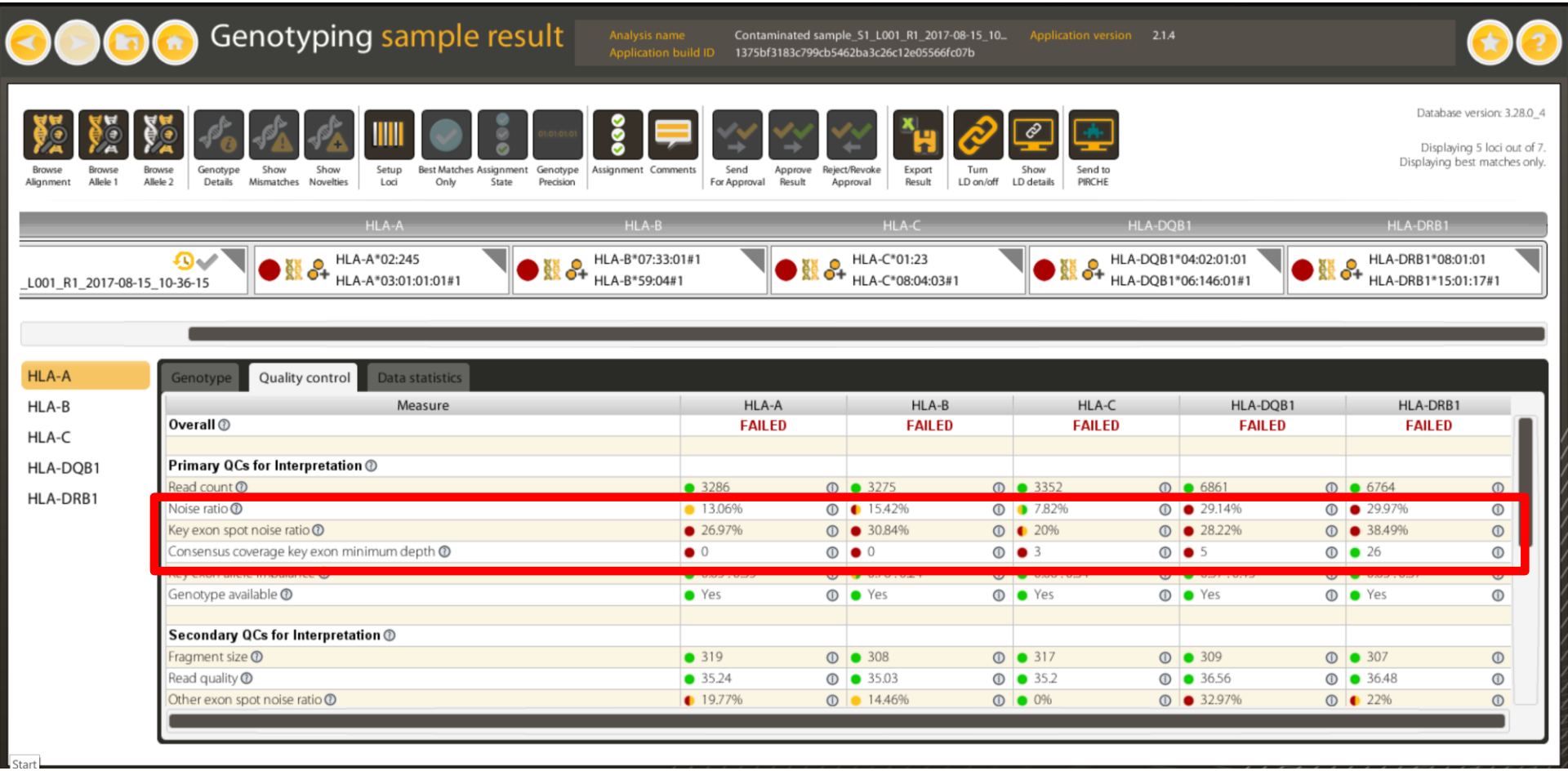
Inspect sequences



Visualize noise reads in the noise track



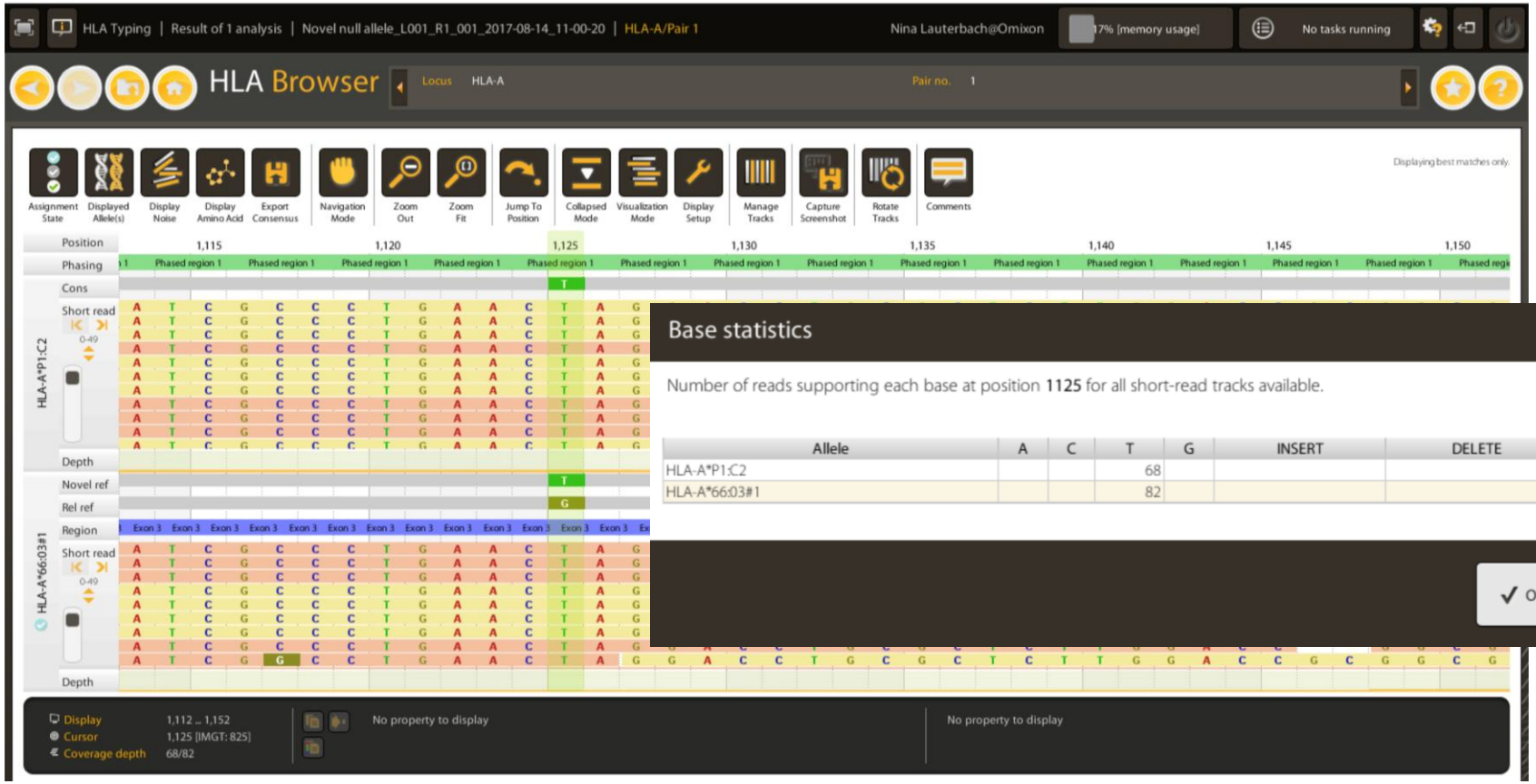
How to be alerted on potential contamination?



Check the noise track…

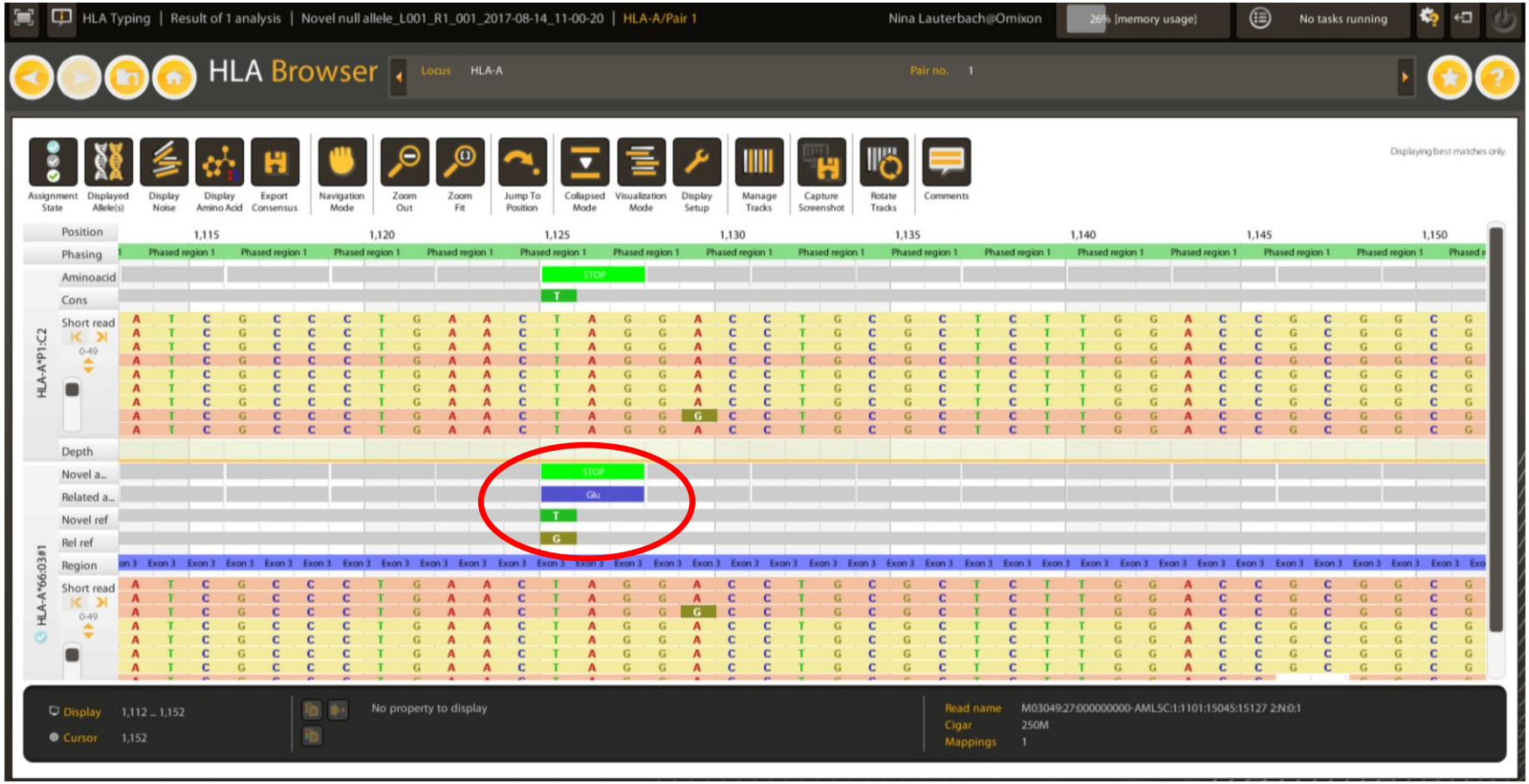
HLA Twin Novel allele confirmation

- You can confirm a novel sequence in the Gene Browser



HLA Twin Novel allele confirmation

- You can confirm a novel sequence in the Gene Browser



Summary :

1. NGS provides us with a large amount of data
   - High resolution typings
   - Complete missing sequences in IMGT database
   - Few to no ambiguities
   - Few to no reflex testing

2. NGS analysis software is a crucial part in obtaining trustworthy typings. Therefore HLA Twin provides:
   - Confidence in typing due to strong algorithms
   - Quality Control metric to alert user for potential issues
   - Tools for troubleshooting

HLA NGS is not just to count alleles…

Important to give a functional meaning to the high resolution data

1. Improve the PIRCHE export function in the software – make it bidirectional

2. Incorporate HLAMatchmaker algorithm

We are constantly learning and improving

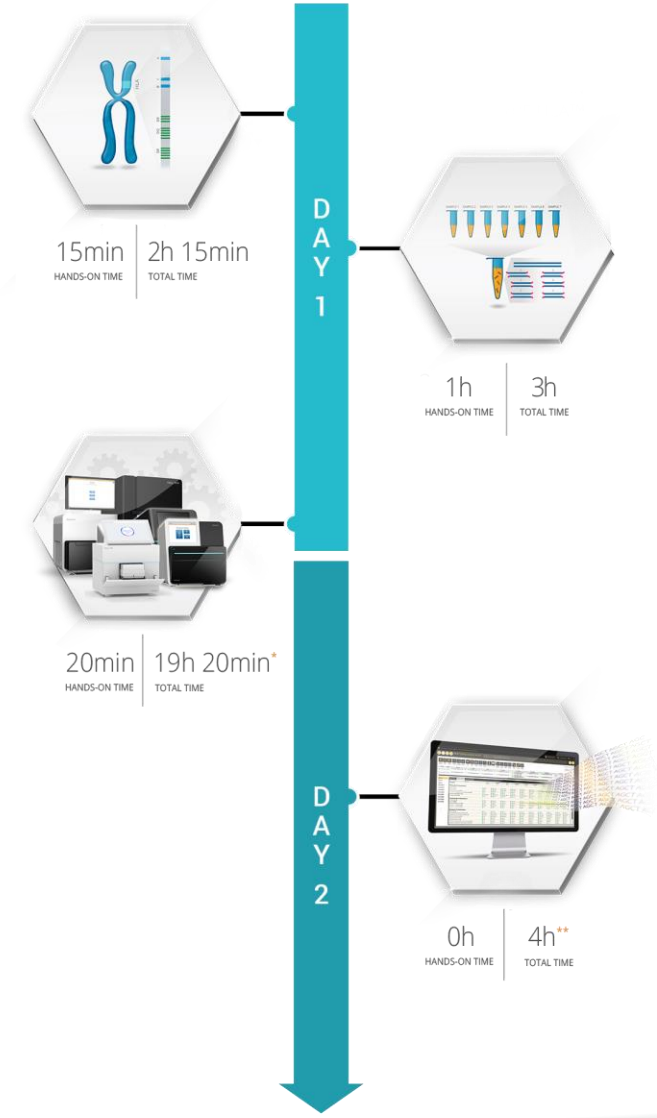

What's
new?

New Product: OmniType

- 11-locus, 1 tube multiplex assay for HLA typing
- Single workday workflow
- Fast PCR
- Efficient and technician-friendly library preparation protocol
- Minimal hands-on time
- Automatable protocol for pre- & post-PCR

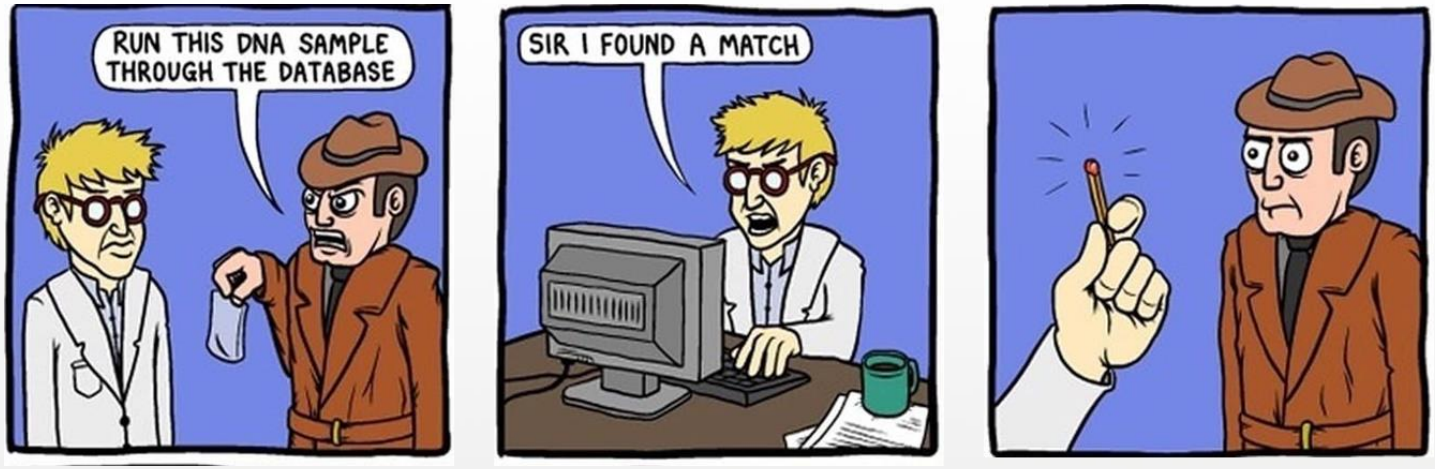
OmniType Workflow

- DNA to library in <7 hours
- ~1 hour hands-on time
- 100 ng total input DNA
- 11 loci, 1 tube
- Fast LR-PCR, ~2 hours
- No amplicon pooling
- Fast library preparation workflow
- Sequence results on Day 2



OmniType Early Access Program (EAP)

- Up to 48 samples FOC
- Up to 96 at special discount for validation
- Free onsite training with an Omixon FAS
- Ability to influence assay & software product development
- Join the EAP now!



Nina Lauterbach, PhD
Senior Field Application Scientist
Nina.lauterbach@omixon.com

Richard Pallin
VP Commercial Operations
Richard.Pallin@omixon.com

www.omixon.com