

The logo for AIBT, consisting of the letters 'AIBT' in a bold, blue, sans-serif font, enclosed within a white rectangular border.

AIBT

Summer School 2015

**Alloreattività e trapianti nell'uomo:
le nuove metodiche di studio
e i trapianti alternativi**

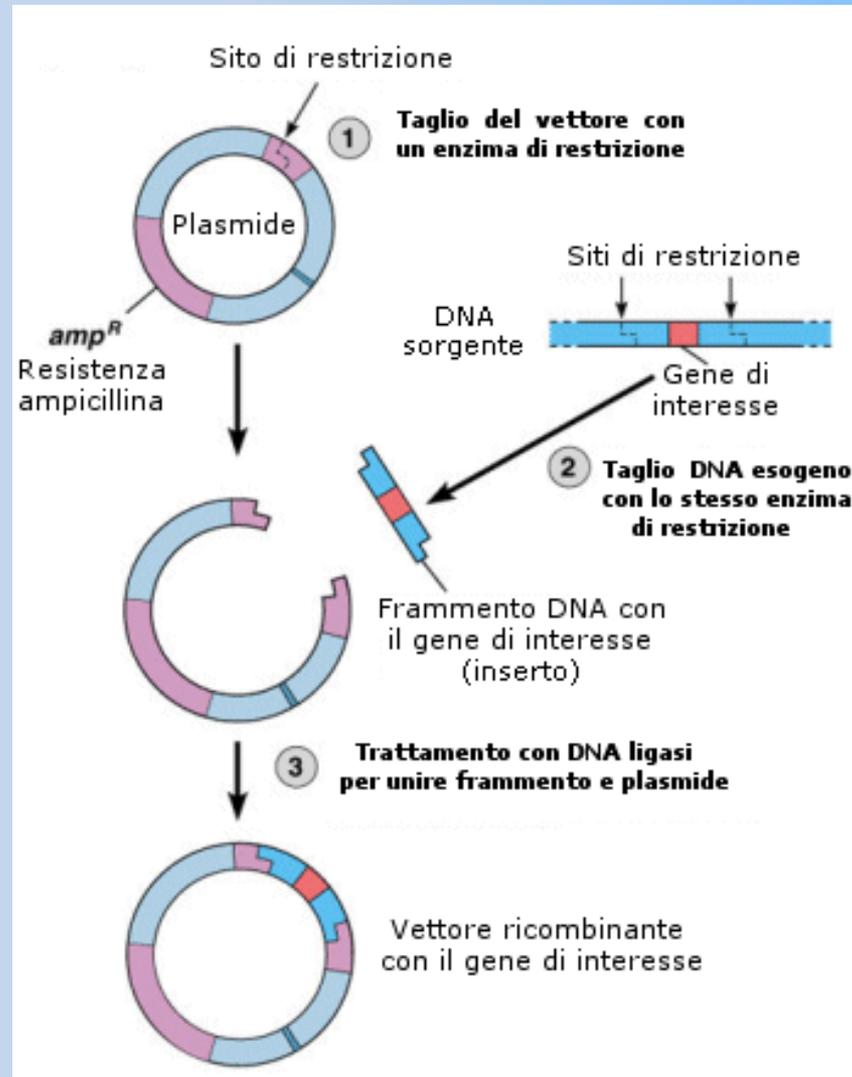
**Applicazioni della Next-Generation
Sequencing con tecnologia Illumina
Roberto Cusano**

04 - 06 giugno 2015 Villaggio Cala la Luna Favignana (TP)

Principi del sequenziamento clonale

- Dai vettori plasmidici agli oligonucleotidi adattatori - **Library Prep**
- Dalle PCR specifiche all'ultra multiplex PCR - **Library Amplification**
- Dalle colonie su piastra ai cluster su flow cell - **Bridge PCR**
- Da 10^3 bp a 10^9 bp - **Deep Sequencing**

La preparazione delle librerie

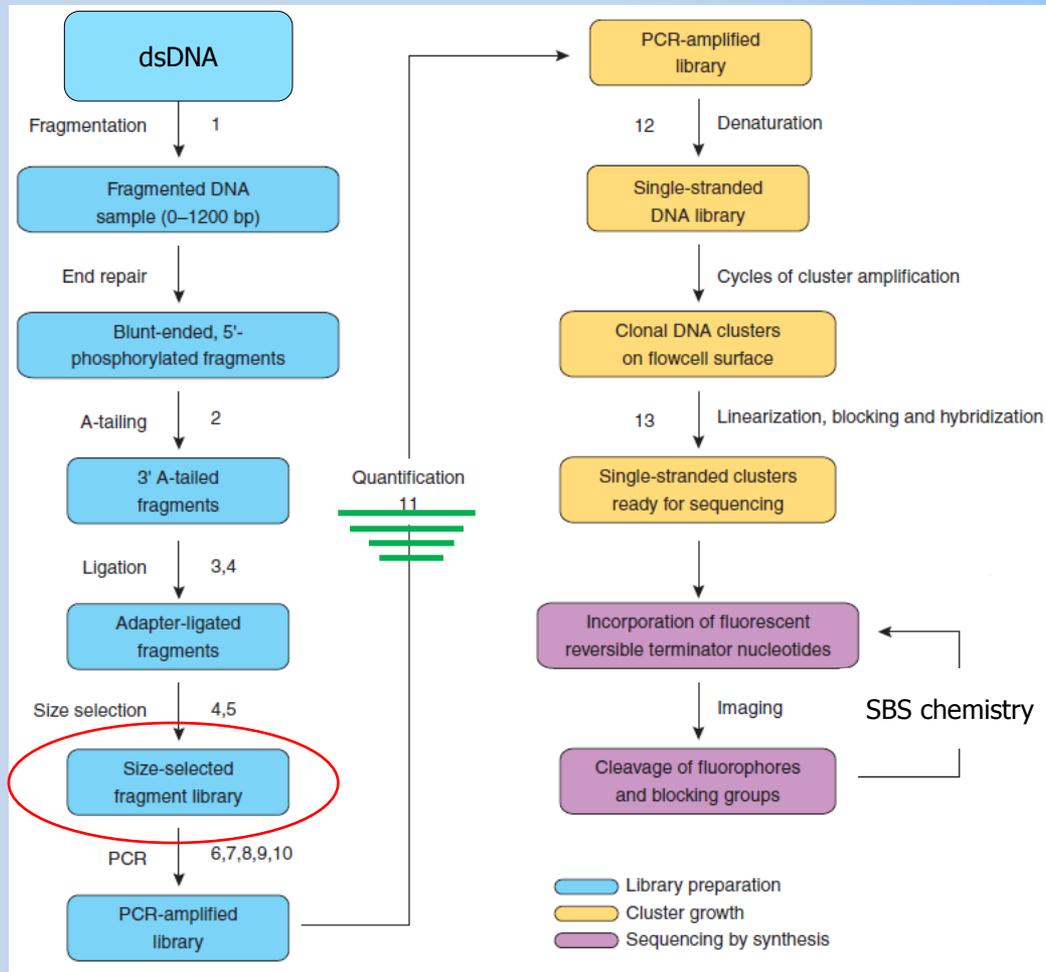


La preparazione delle librerie

Genomic DNA, dscDNA, LR-PCR, ChIP, ecc...

Metodo classico

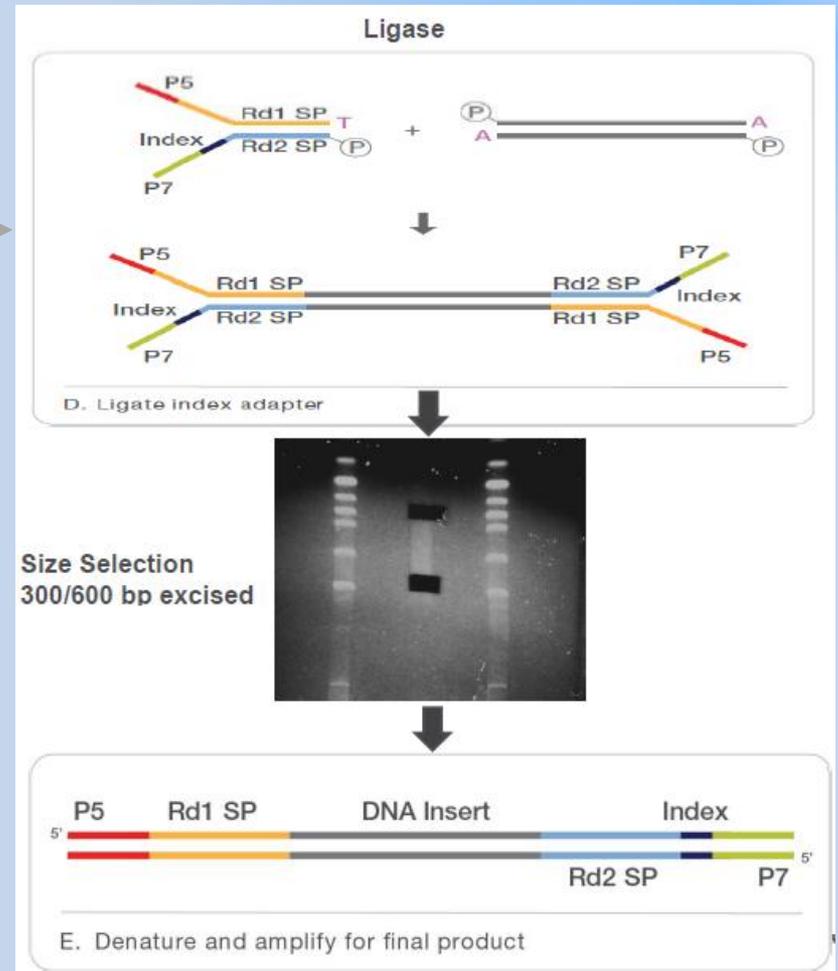
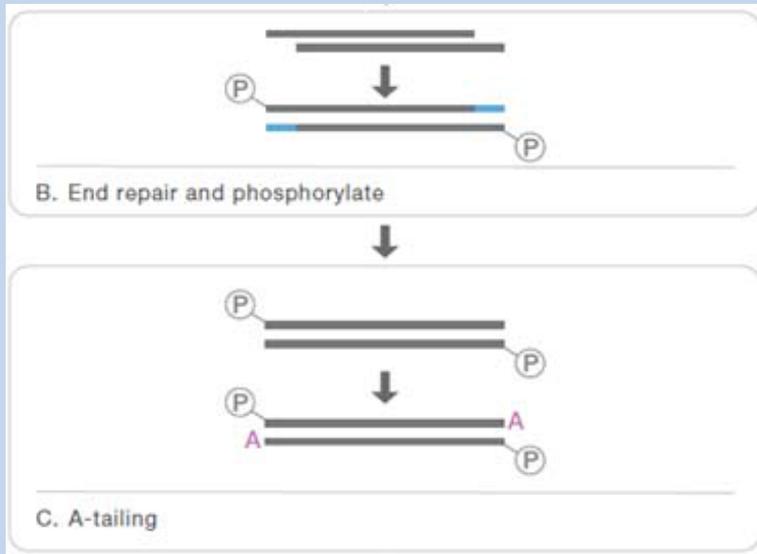
-Mappabilità
-Clustering



La preparazione delle librerie

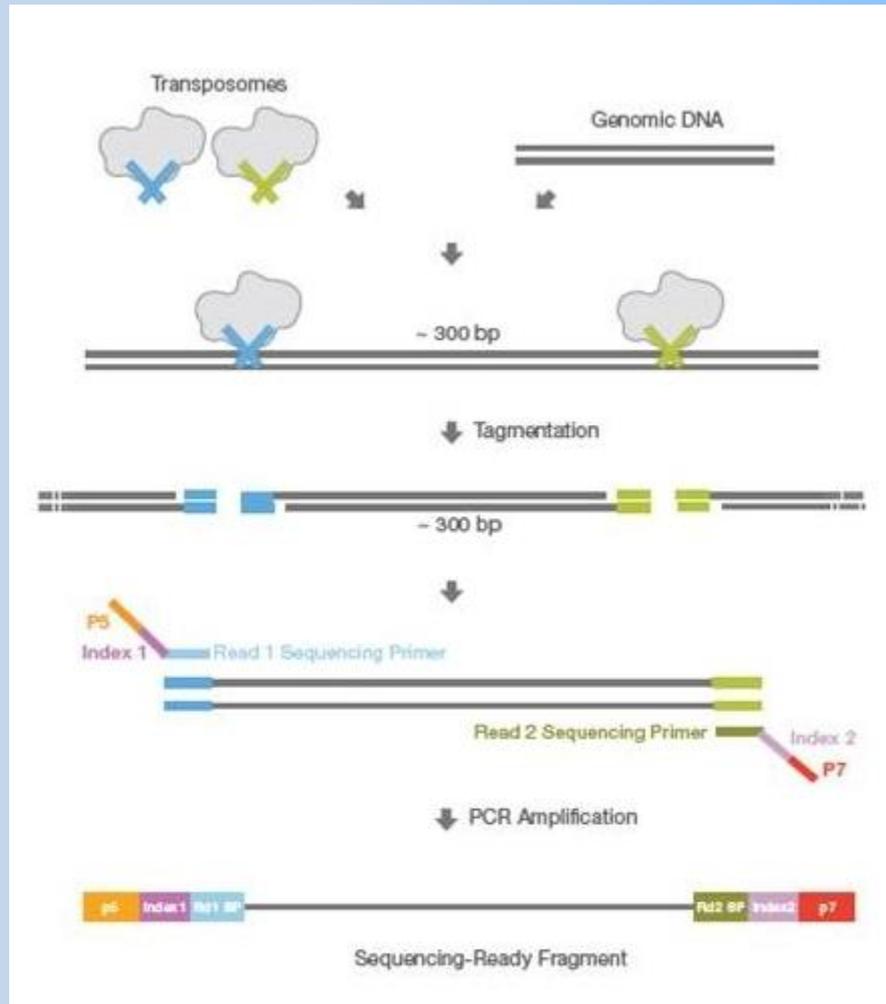
Size selection da gel di agaroso

dsDNA fragments



La preparazione delle librerie

Metodo della tagmentazione

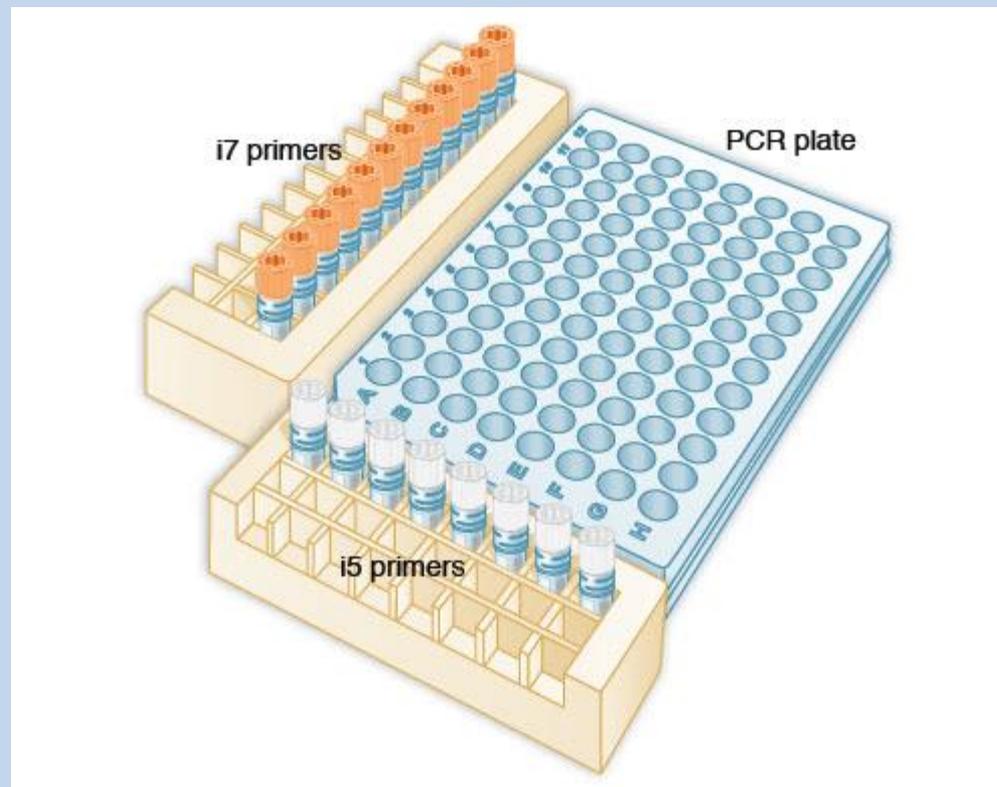


Nextera chemistry simultaneously fragments and tags DNA in a single step. A simple PCR amplification then appends sequencing adapters and sample indices to each fragment.

La preparazione delle librerie

Sistema dual-index

8 index primer i5 x 12 index primer i7 = 96 combinazioni
4 diversi set di primer = 384 combinazioni su una flow cell



La preparazione delle librerie

Strategie di indexing

➤ Individual tagging

Ogni campione 'consuma' una combinazione di i5 + i7 primer e una reazione di tagmentazione.

Vantaggi: minor costo per campione e maggiore processività

Svantaggi: maggiore probabilità di errori di allineamento

➤ Locus tagging

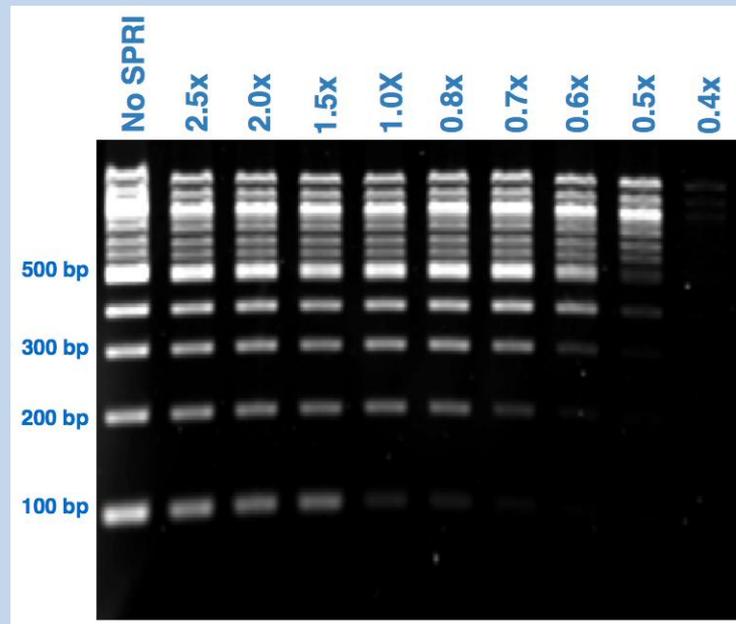
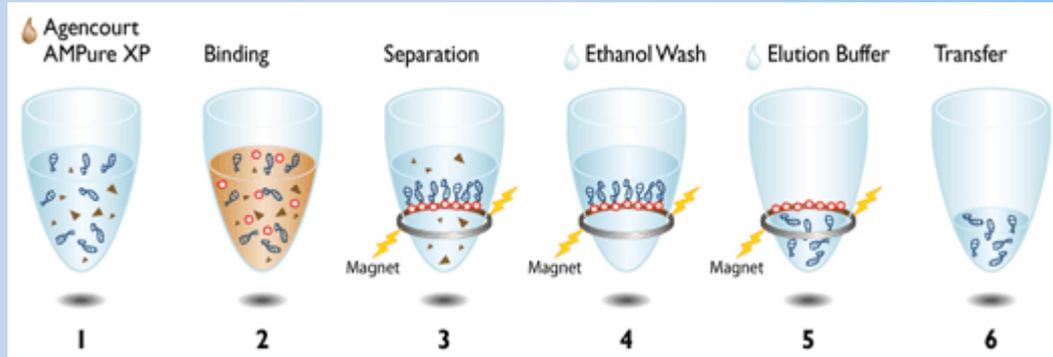
Ogni gene 'consuma' una combinazione di i5 + i7 primer e una reazione di tagmentazione.

Vantaggi: maggiore specificità nell'allineamento

Svantaggi: maggiore costo per campione e bassa processività

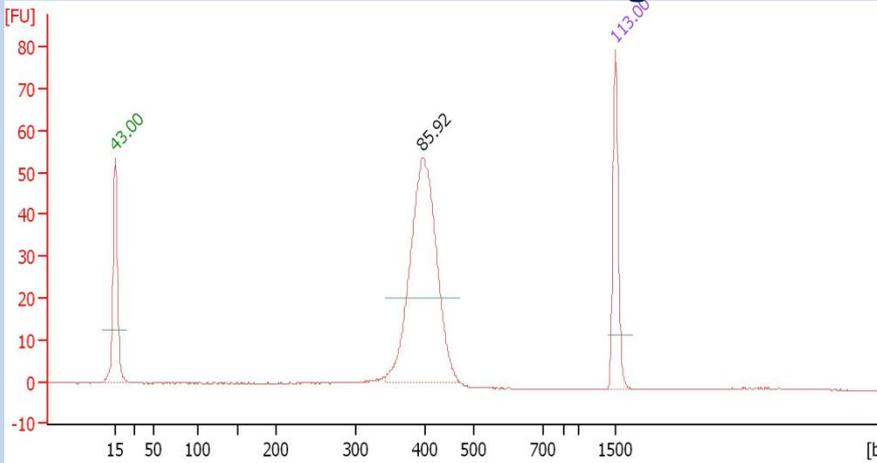
La preparazione delle librerie

Size selection con biglie magnetiche



La preparazione delle librerie

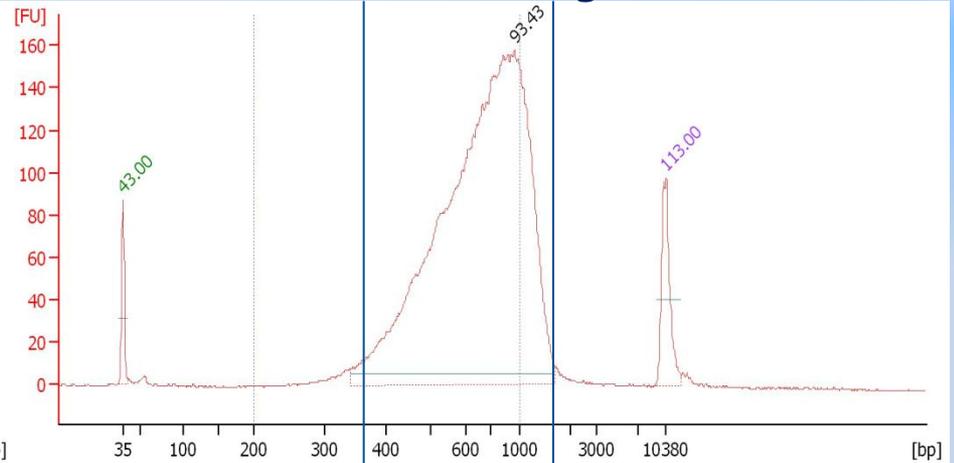
Selezione da gel



Peak	Size [bp]	Conc. [ng/μl]	Molarity [nmol/l]
1	15	4.20	424.2
2	396	8.88	34.0
3	1,500	2.10	2.1

Observations
Lower Marker
Upper Marker

Selezione con biglie



Peak	Size [bp]	Conc. [pg/μl]	Molarity [pmol/l]	Observations
1	35	125.00	5,411.3	Lower Marker
2	940	1,790.38	2,886.1	
3	10,380	75.00	10.9	Upper Marker

2° step, beads 0,6X
lega frammenti più piccoli

1° step, beads 0,5X
lega frammenti grandi



Discard supernatant
Keep beads

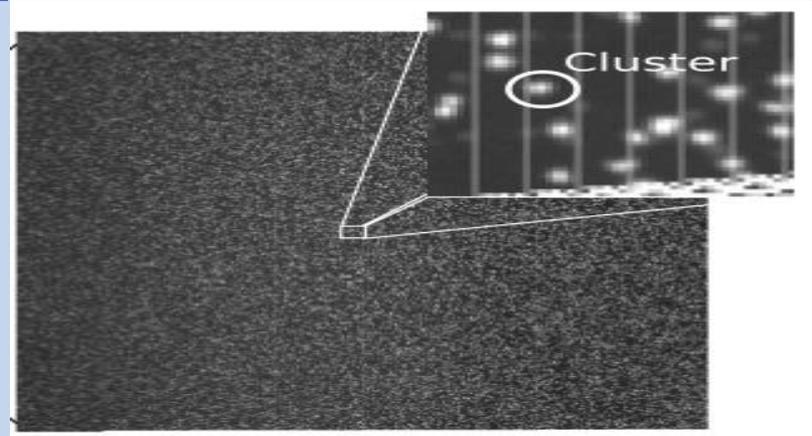


Keep supernatant
Discard beads

La generazione dei cluster

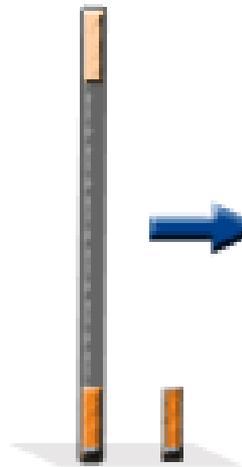


La generazione dei cluster

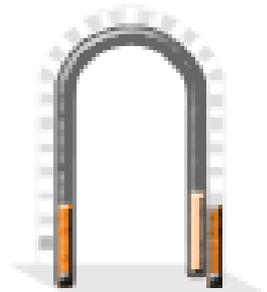


La generazione dei cluster

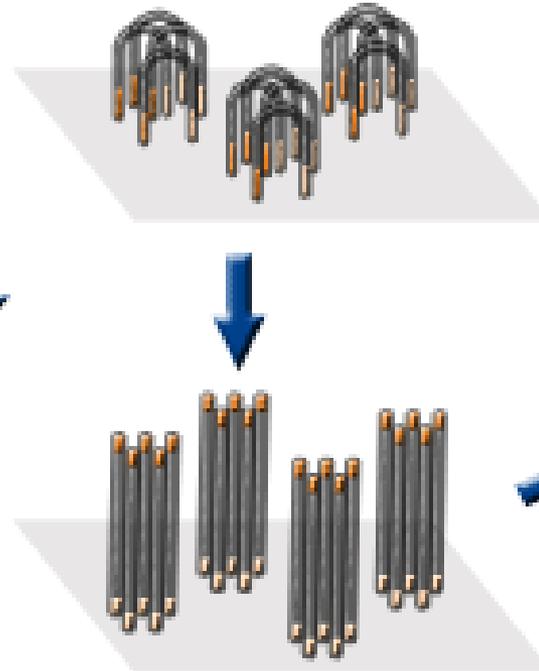
1. Attach DNA to flow cell



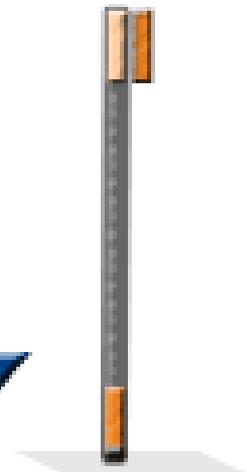
2. Perform bridge amplification



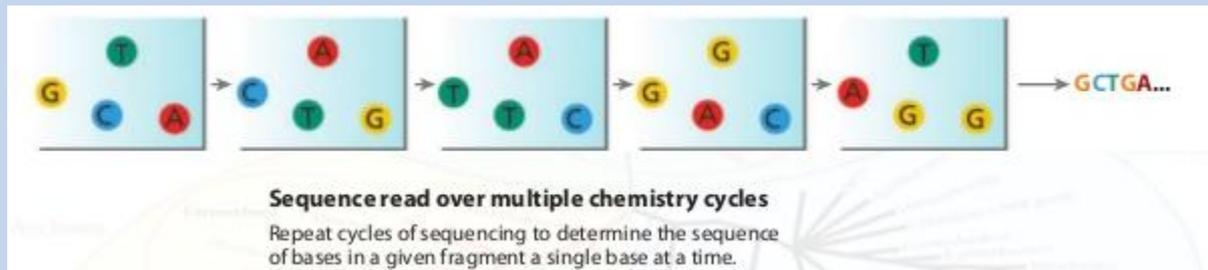
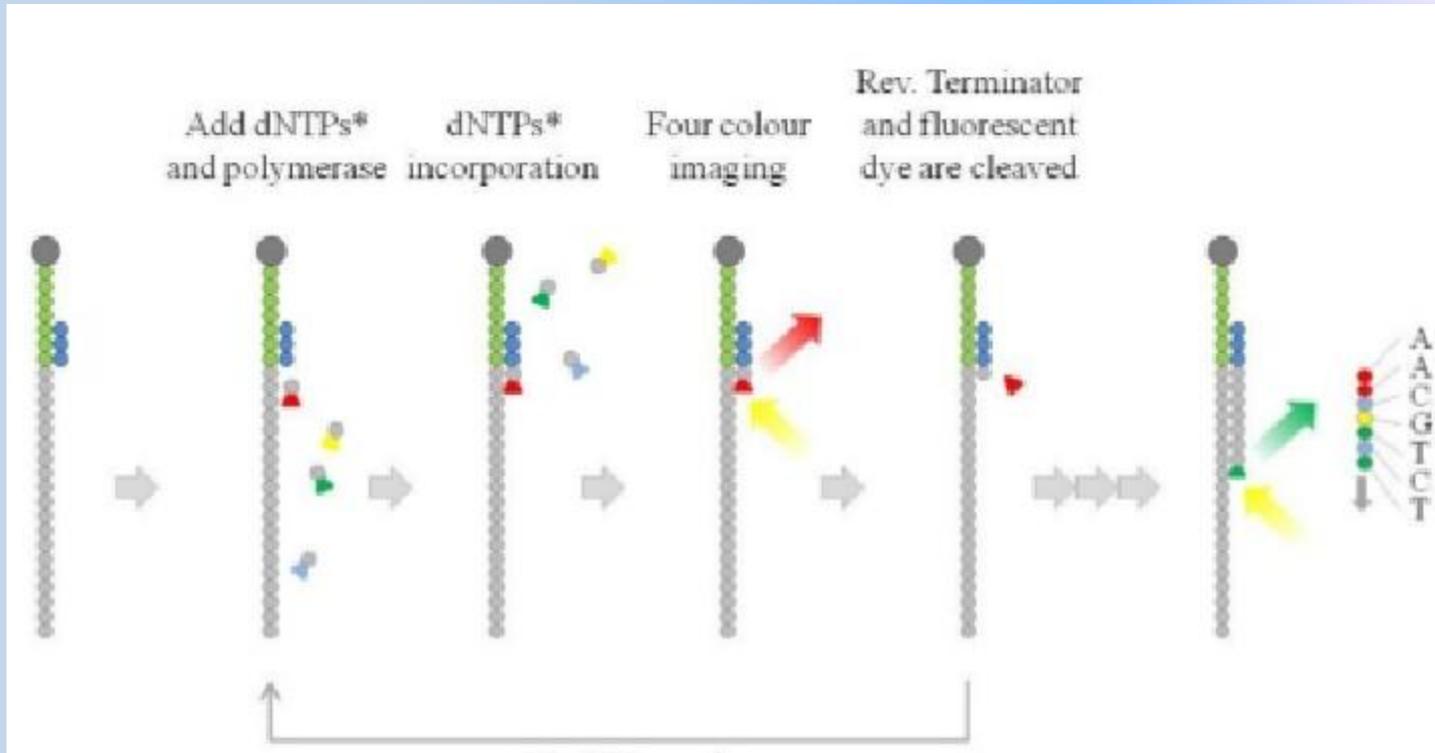
3. Generate clusters



4. Anneal sequencing primer



Chimica SBS



Sequenziamento Massivo e Parallelo

La massima resa di una flow cell è data dal numero di cluster e dalla lunghezza delle read.

Per esempio:

MiSeq Flow Cell v3 600 cicli:

25MCluster in Paired-End o 50MReads (300+300) = **15Gb**

Per un target di 45Kb (A-B-C-DRB1-DPB1-DQB1 full length) e 192 campioni sequenziati si ottiene una copertura media teorica di 1700X (260000 Read per campione)

Ogni cluster produce 4 read:

- Read 1, sequenza forward dell'inserto (301)
- Read 2, sequenza reverse dell'inserto (301)
- Index read 1 (8)
- Index read2 (8)

Sequenziamento Massivo e Parallelo

Per ogni campione vengono generati due fastq file, read1 e read2

```
1 @M03256:2:000000000-ABR39:1:1110:24607:17535 1:N:0:72 _____ numero della read
2 GCCCTACTCCGTCCCTTTCTCTATCCACATTGCTTTAAATCATATTTTCTCTCAAGGTGTACAAGGATGATAAATAGGTGCCAAGTG
3 +
4 CC@CCFGGGGFDGFGGGGFGFCFGFFFC, FFGGFFGFAEEF@FF9CFFGFFGC<C8ECFFGGFF, , ;<CEFFFGG<, @@AFEAFGF

@M03256:2:000000000-ABR39:1:1110:24607:17535 2:N:0:72
ATCTTTATAGTGGGGACCCATTAGATTTGAGAGATCTTGTGAAAAATTATGTTTGGCTCTTCATAGCTTGAAATTGACATGCATTGTC
+
CCC@, CCF, <FECC7; @, 8@A<FGGGGFAFGGGFG, EEFEFFFG, @<EAFFGGG, :<FF@&#x27;&#x27;FFDFGGGG?AFFGC<F, 9CF, , AA@F
```

- 1 - identifica e descrive la sequenza
- 2 - sequenza in nucleotidi
- 3 - riga *opzionale*
- 4 - codifica i valori di qualità delle singole basi della sequenza

Un fastq file che contiene 140 KReads pesa 80MB
Una Run su MiSeq produce circa 40 GB in fastq files

Principali Applicazioni

Multiplex
Exon Target
PCR

Long Range
PCR

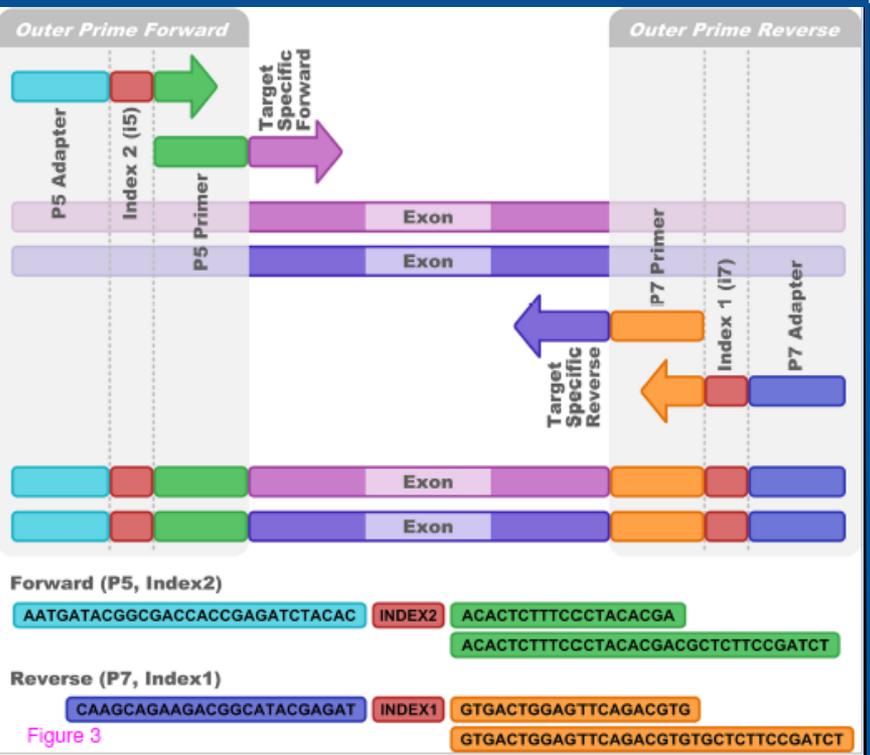
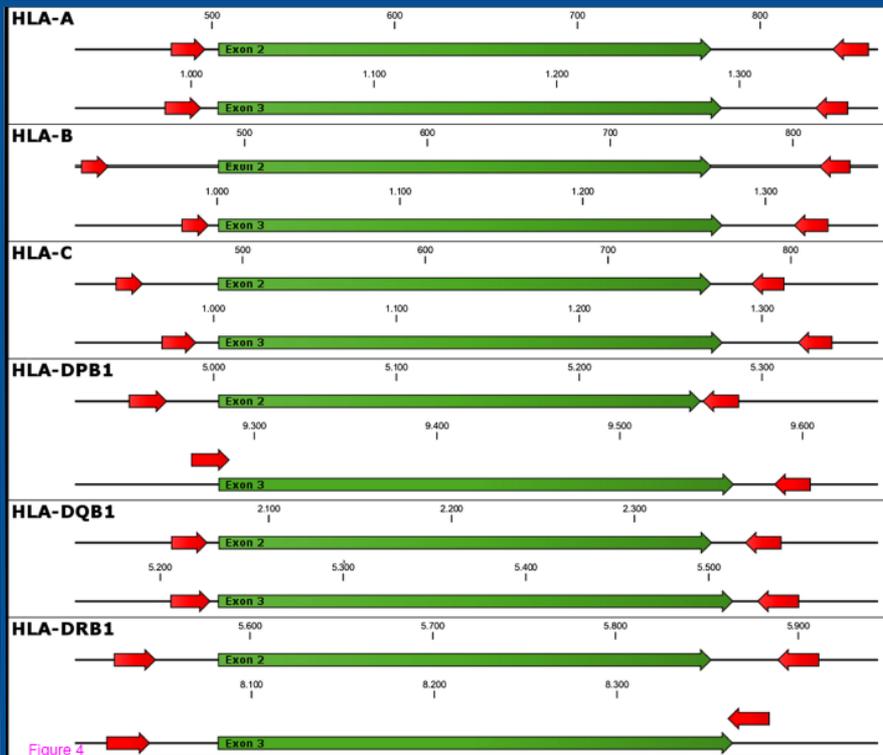
Genomic
Target
Capture

Whole
Exome

Transcripts
Target
Capture

Whole
Genome

High-throughput Multiplex Exon Target PCR



"Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing"
 Lange V. - *BMC Genomics*, 2014 (DKMS)

High-throughput Multiplex Exon Target PCR

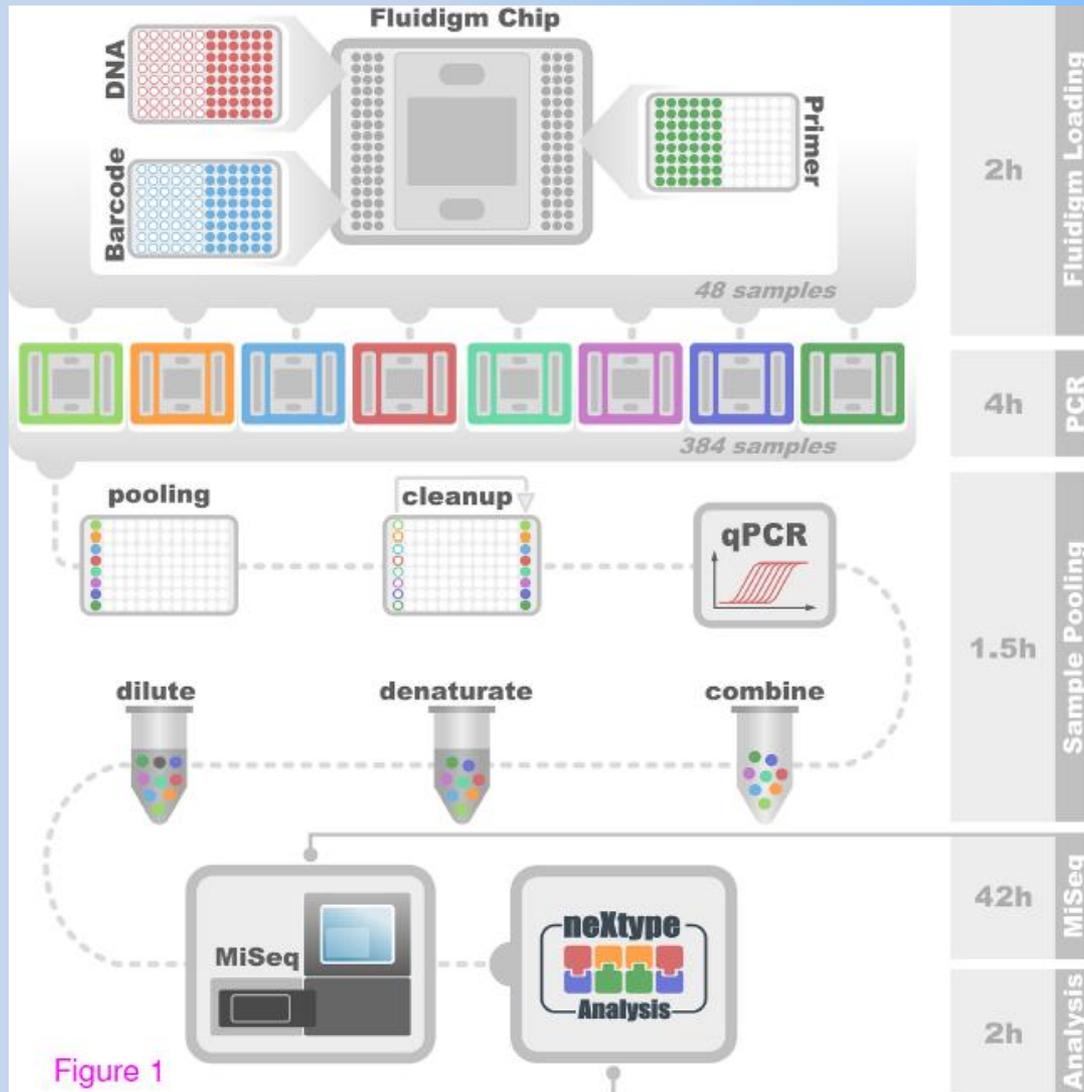
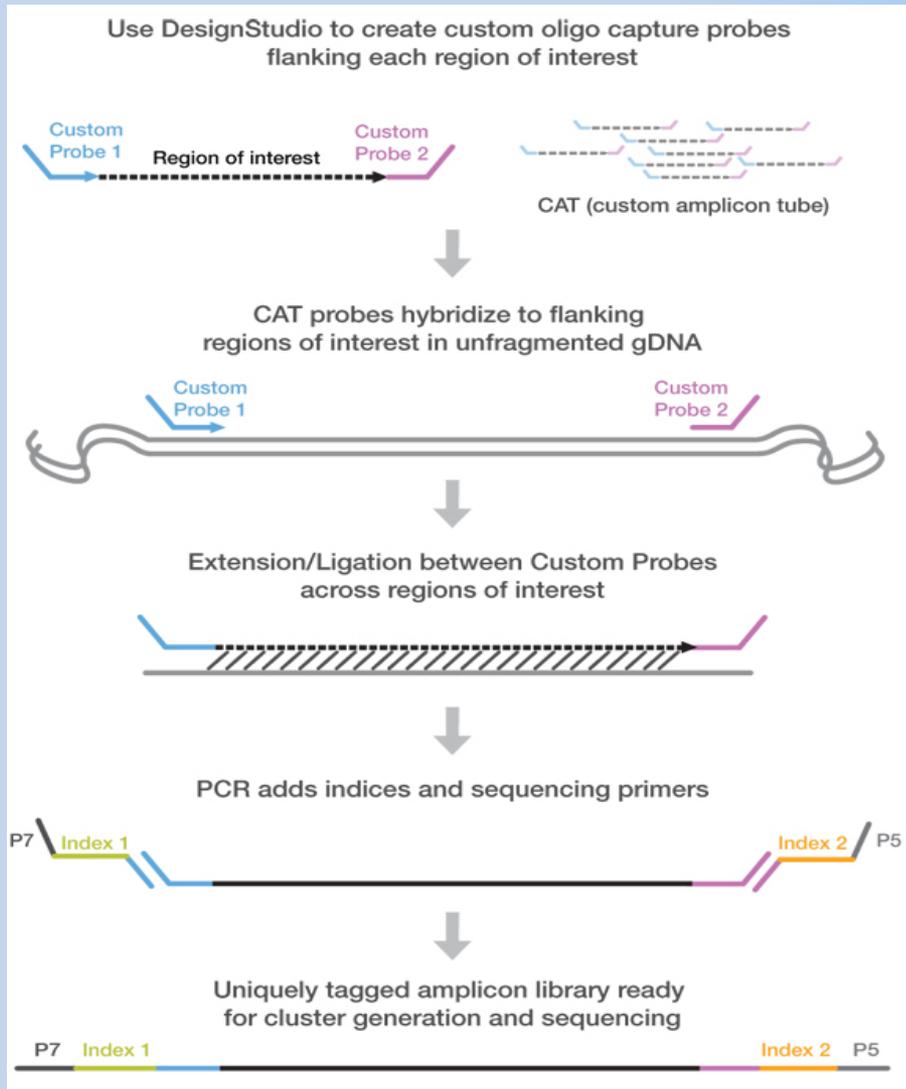


Figure 1

Illumina TruSeq Custom Amplicon



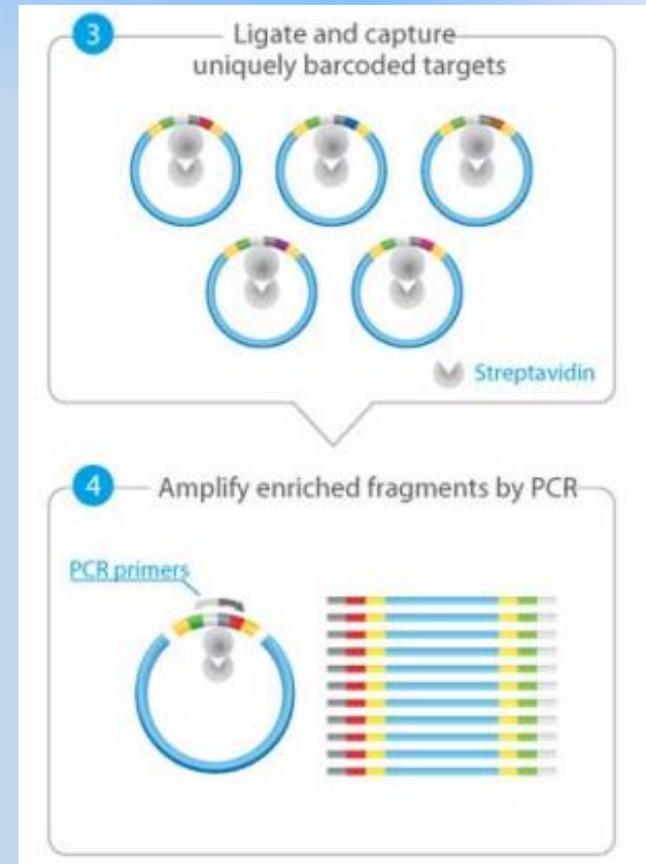
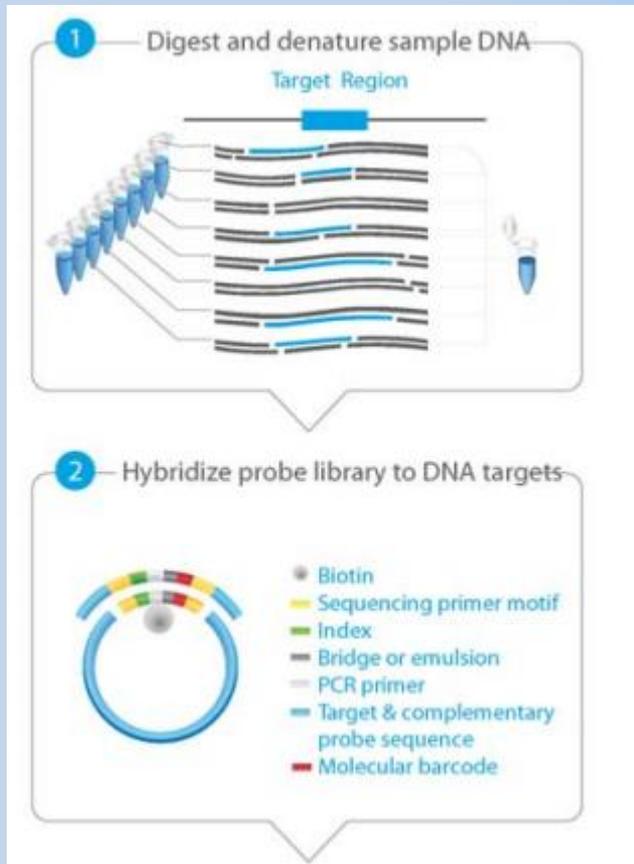
Fino a 1,536 ampliconi per un target di ~2-650 Kb

In singola provetta con un sistema di ligazione amplificazione

Diverse grandezze degli ampliconi in base al tipo di target e del successivo sequenziamento

L'analisi può essere completamente automatizzata

Agilent HaloPlex Target Enrichment



-Diversi disegni possibili con target da 500Kb per 20.000 probe fino a 5Mb

-Privo di PCR bias ed ha un sistema di barcode molecolare che permette di riconoscere i duplicati di PCR

High-throughput Multiplex Exon Target PCR

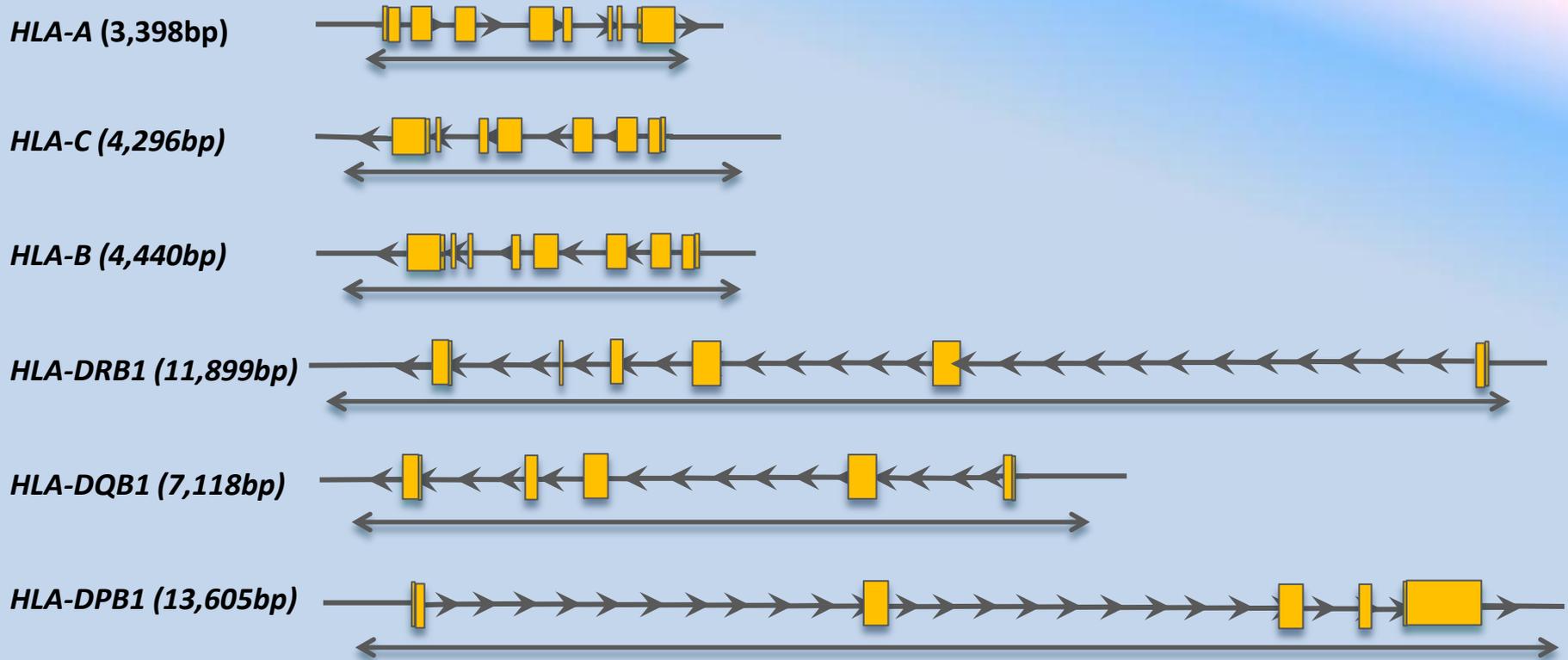
Vantaggi:

- Non sono necessari step di preparazione delle librerie perché i primer di PCR includono le sequenze necessarie per l'ibridazione sulla flow cell e per il sequenziamento
- Il metodo è completamente automatizzabile

Svantaggi:

- E' difficile disegnare primer che comprendano tutti gli alleli per tutti gli esoni da analizzare e sono necessari pool di primer molto complessi
- Non si hanno informazioni sulla fase tra varianti in esoni diversi

Long Range PCR



"Phase-defined complete sequencing of the HLA genes by next-generation sequencing"
Hosomichi K. - *BMC Genomics*, 2013

"A Bead-based Normalization for Uniform Sequencing depth (BeNUS) protocol for multi-samples sequencing exemplified by *HLA-B*" Hosomichi K. - *BMC Genomics* 2014

Long Range PCR

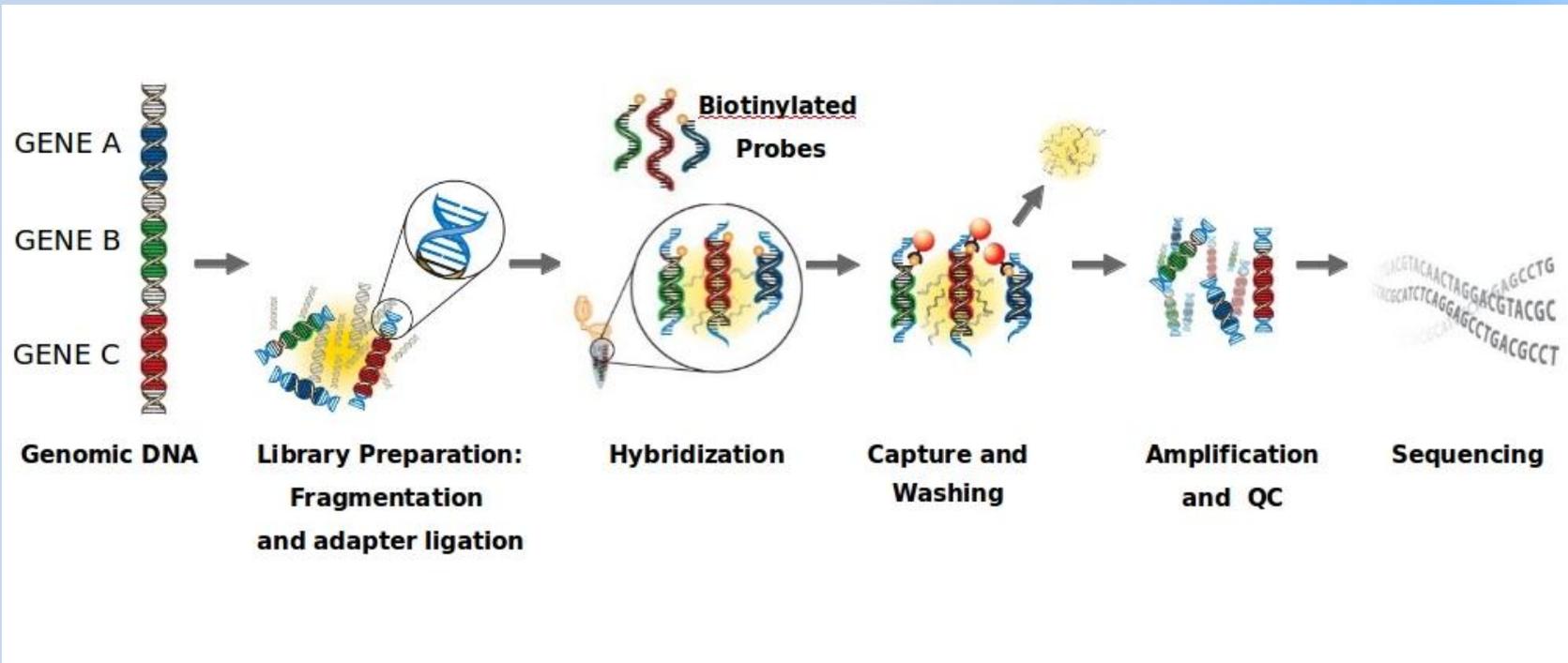
Vantaggi:

- Possono essere caratterizzati con maggiore dettaglio i geni
- Si hanno informazioni sulla fase tra varianti in esoni diversi
- Possono essere aggiunti ampliconi in ogni momento
- Si possono fare multiplex PCR
- Molti step di preparazione possono essere automatizzati

Svantaggi:

- La preparazione prevede molti step ed è necessaria una elevata precisione per avere risultati consistenti
- Elevato rischio di inquinamento perché sono coinvolti due diverse amplificazioni
- L'amplificazione preferenziale e la produzione di artefatti tipici della PCR (errori della polimerasi, prodotti chimerici, aspecifici)

Genomic Target Capture



"Development of a high-resolution NGS-based HLA-typing and analysis pipeline"
Wittig M. - *Nucleic Acids Research*, 2015

Genomic Target Capture

Workflow

- ✓ Preparazione delle librerie a partire da DNA genomico
- ✓ Prima PCR di indexing
- ✓ Produzione dei pool di campioni
- ✓ Cattura per ibridazione dei geni d'interesse
- ✓ Amplificazione di arricchimento
- ✓ Dosaggio e QC delle librerie
- ✓ Produzione di un pool con l'intero set di campioni
- ✓ Corsa su MiSeq

Genomic Target Capture

Vantaggi:

- E' una metodica molto consolidata (Exome Seq)
- Non si ha lo sbilanciamento allelico tipico della PCR
- Non ci sono limiti alla grandezza delle regioni e al numero di target
- Ha un workflow rapido perché è in singolo tubo
- Molti step di preparazione possono essere automatizzati
- Si possono facilmente aggiungere nuovi geni

Svantaggi:

- Per coprire tutti gli alleli è necessario un grande numero di sonde
- La cattura di regioni più grandi rispetto all'exome potrebbe essere un problema
- La ricostruzione degli aplotipi potrebbe essere più complicata

Whole Genome

Whole Exome

Whole Transcriptome

Transcripts Target Capture

"HLA Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data". Major E. - PLOS, 2013

"Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads". Bai Y - *BMC Genomics* 2014

"HLA typing from RNA-Seq sequence reads".
Boegel S. - *Genome Medicine*, 2012

La nostra esperienza

Messa a punto di un protocollo home-made per l'analisi dei geni HLA

- 7 Prodotti di LongRange PCR dal 5' al 3' UTR per i loci: A, B, C, DRB1, DPB1, DQB1, G per un target di ~ 50Kb (primer Hosomichi modificati)
- Preparazione delle librerie con kit Nextera XT modificato e individual tag
- Selezione dei frammenti mediante doppio taglio con AmpureXP Beads
- Normalizzazione delle librerie con Nextera XT normalization module
- Sequenziamento su Illumina MiSeq con Flow Cell v3 300+300 cicli
- Comparazione di diversi software per l'analisi dei dati

Workflow

(tempi riferiti a una plate da 96 campioni)

1. Amplificazione O/N dei loci d'interesse

1° giorno

2. Dosaggio dei frammenti e produzione di un pool di prodotti equimolari

3. Dosaggio fluorimetrico e diluizione intermedia dei pool

2° giorno

4. Dosaggio fluorimetrico e diluizione di lavoro dei pool

5. Tagmentazione e purificazione

6. Indexing PCR

7. Doppia purificazione per la selezione dei frammenti

8. Normalizzazione delle librerie

3° giorno

9. Produzione del pool di librerie equimolari

10. Sequenziamento su Illumina MiSeq con Flow Cell v3 301+301 cicli

56h

11. Analisi dei dati

Analisi delle sequenze

Software per l'analisi di geni HLA attualmente commercializzati

Assign MPS - Conexio Genomics

HLA target - Omixon

NGSEngine - GenDx

SeqNext-HLA - JSI

Software Open-Source

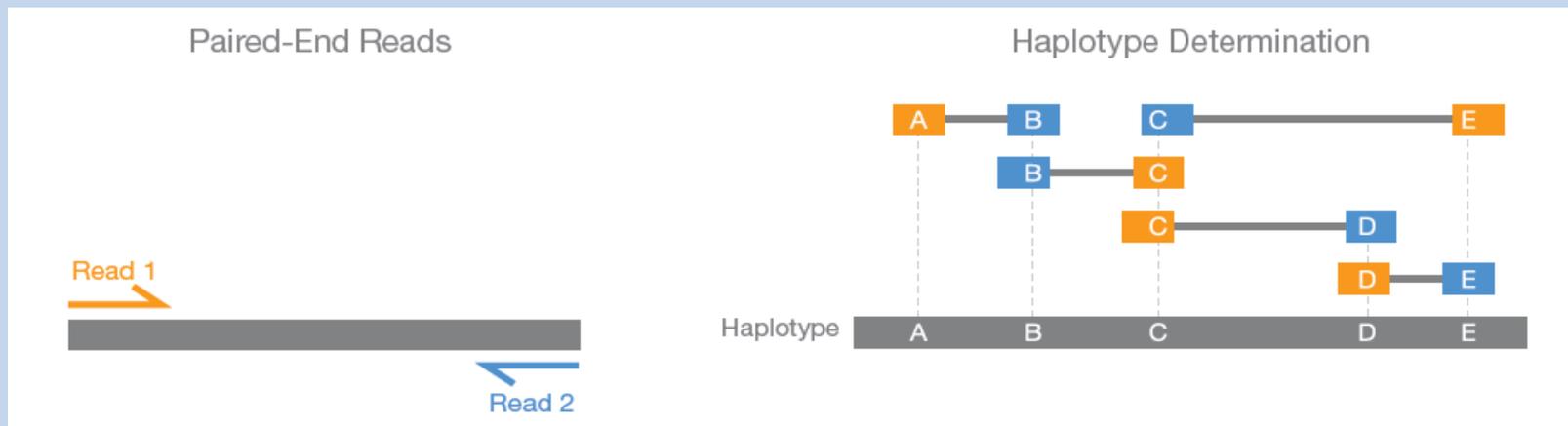
HLAssign - <http://www.ikmb.uni-kiel.de/>

"Development of a high-resolution NGS-based HLA-typing and analysis pipeline" Wittig M. - *Nucleic Acids Research*, 2015

Analisi delle sequenze

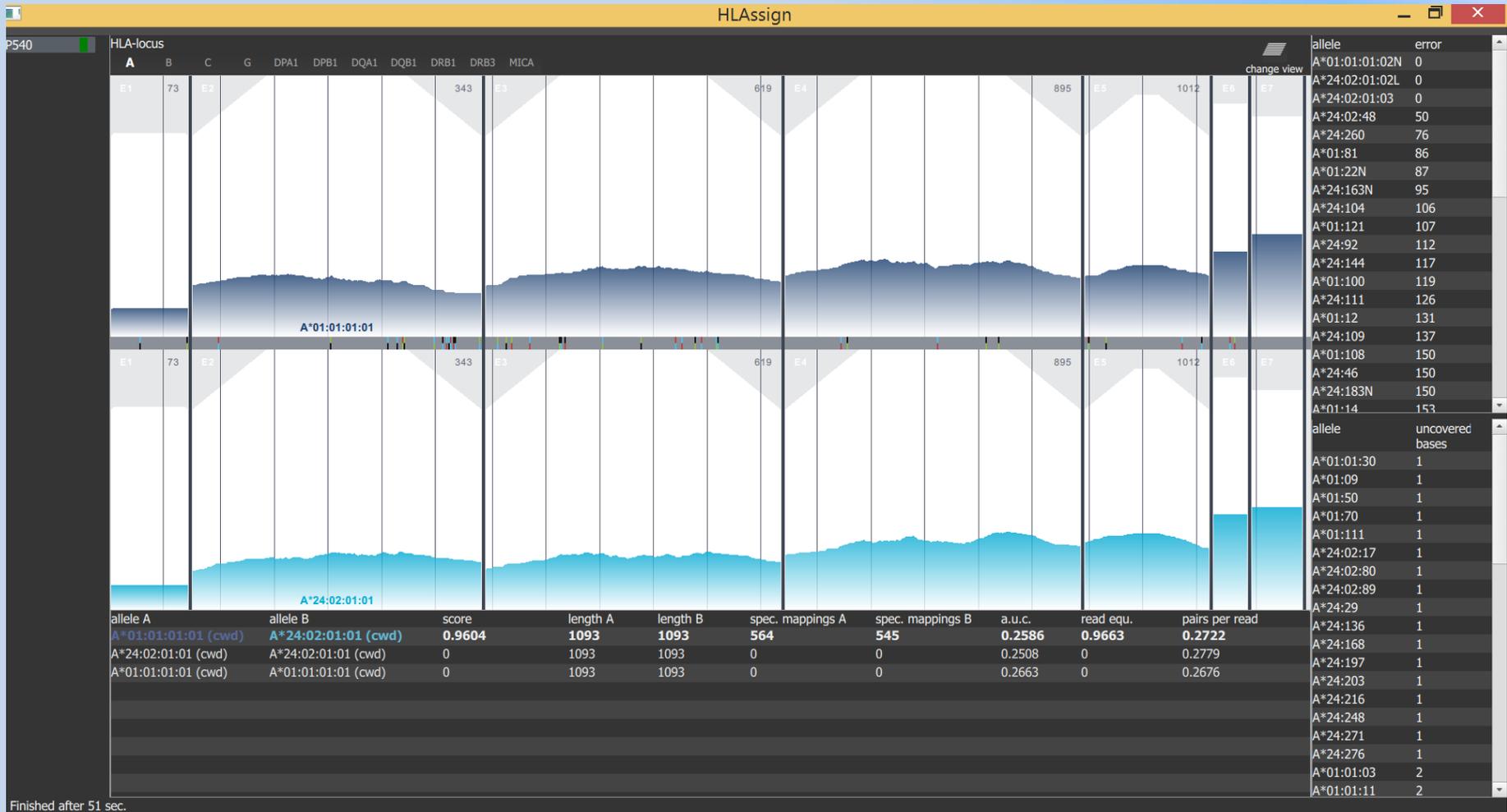
In generale l'analisi si basa su diversi algoritmi:

- Alignment - in riferimento a una libreria di geni e alleli (IMGT/HLA Database) vengono ordinate le read e identificate le varianti
- Haplotype Phasing - Le eterozigosi vengono valutate per determinare la loro relazione cis/trans
- Genotype Determination - Gli aplotipi ottenuti vengono paragonati alla libreria di alleli per assegnare un grado di corrispondenza



HLAssign

<http://www.ikmb.uni-kiel.de/>



Finished after 51 sec.

Assign Conexio Genomics

A3 BaseCalls - Assign

Home

Import: **conexio** | Edit: | Type: **Genotyping** | Codons | Fixed Width | Collapsed | Grouping: **All Alleles** | Show: **Minir**

Update: | Format: **Excel** | All Layers | Elastic Scroll | Sort By: **Sample** | CWD Set: **[None]** | Fields: **All** | Annotation: **[None]**

Edit: | Reports: | Unaligned | Filtered | **ACCEPT** | 1039 | 99 | Go

Data: | Settings: | References: | Reports: | Options: | No Offset | Master | BCS | Edits | MM | Var | Auto | Rare | Het

BaseCalls

5' UTR | Ex | Intr | Exon 2 | Intron 2 | Exon 3 | Intron 3 | Exon 4 | Int | Exo | Intron 5 | Intr | Intron | 3' UTR

IMGT/A 3.15.0.0... .1011.....1021.....1031...1041.....1051.....1061. P1629P540cut B A

Base 1039 (1039) GGGCCAGGTTCTCACACCVTCCAGAKRATGTDGTGGCTGCGACGTGGGGYCGGACKGGCC Start: 164 (164) 5' UTR 164

Exon ... C A 2 1 R Stop: 3524 (3524) 3' UTR 300

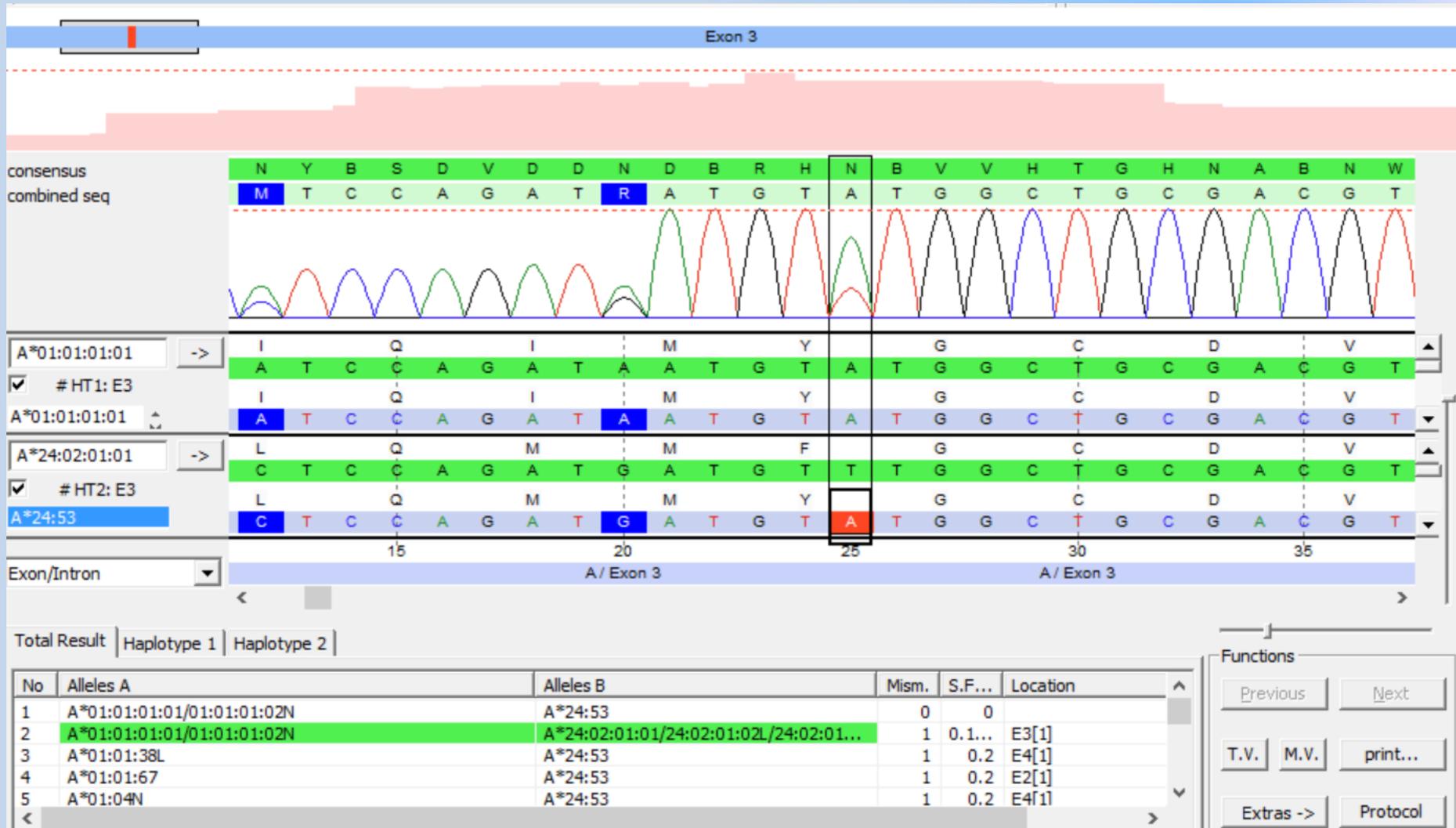
	Allele 1	Allele 2	CORE	EXONS	PHASE1	PHASE2	Differen
P1629P540cut B*	A*****A.....A.....A.....A.....C.....	A*01:01:01:01	0	0	0	0	
B*****C.....G.....T.....T.....	A*01:01:01:01	A*24:02:01:02L	0	0	0	0	Intron 2
C GGGCCAGGTTCTCACACCVTCCAGAKRATGTDGTGGCTGCGACGTGGGGYCGGACKGGCC	A*01:01:01:01	A*24:02:01:03	0	0	0	0	Intron 3
...	A*01:01:01:02N	A*24:02:01:01	0	0	0	0	Intron 2
...	A*01:01:01:02N	A*24:02:01:02L	0	0	0	0	Intron 2
H	A*01:01:01:02N	A*24:02:01:03	0	0	0	0	Intron 2
	A*01:01:26	A*24:02:02	0	0	1	1	Exon 3
	A*01:25	A*24:87	0	0	1	1	Exon 3
	A*01:26	A*24:136	0	0	1	1	Exon 3
	A*01:69:01	A*24:243	0	0	1	1	Exon 2
	A*01:89	A*24:120	0	0	1	1	Exon 3
	A*01:98	A*24:160	0	0	1	1	Exon 3
	A*01:100	A*24:260	0	0	1	1	Exon 2
	A*01:126	A*24:59	0	0	1	1	Exon 3
	A*01:14	A*24:46	0	0	2	2	Exon 3
	A*01:21	A*24:229	0	0	2	2	Exon 3
	A*01:30	A*24:17	0	0	3	3	Exon 3
	A*01:95	A*24:04	0	0	7	7	Exon 2
	A*01:01:01:01	A*24:02:40	0	1	0	1	Exon 5
	A*01:01:01:02N	A*24:02:40	0	1	0	1	Intron 2
	A*01:01:01:01	A*24:02:02	1	0	--	--	
	A*01:01:01:01	A*24:02:03Q	1	0	--	--	
	A*01:01:01:01	A*24:02:04	1	0	--	--	

10% (100)

1% (10)

SeqPilon-NextSeqHLA

JSI Medical Systems



Risultati

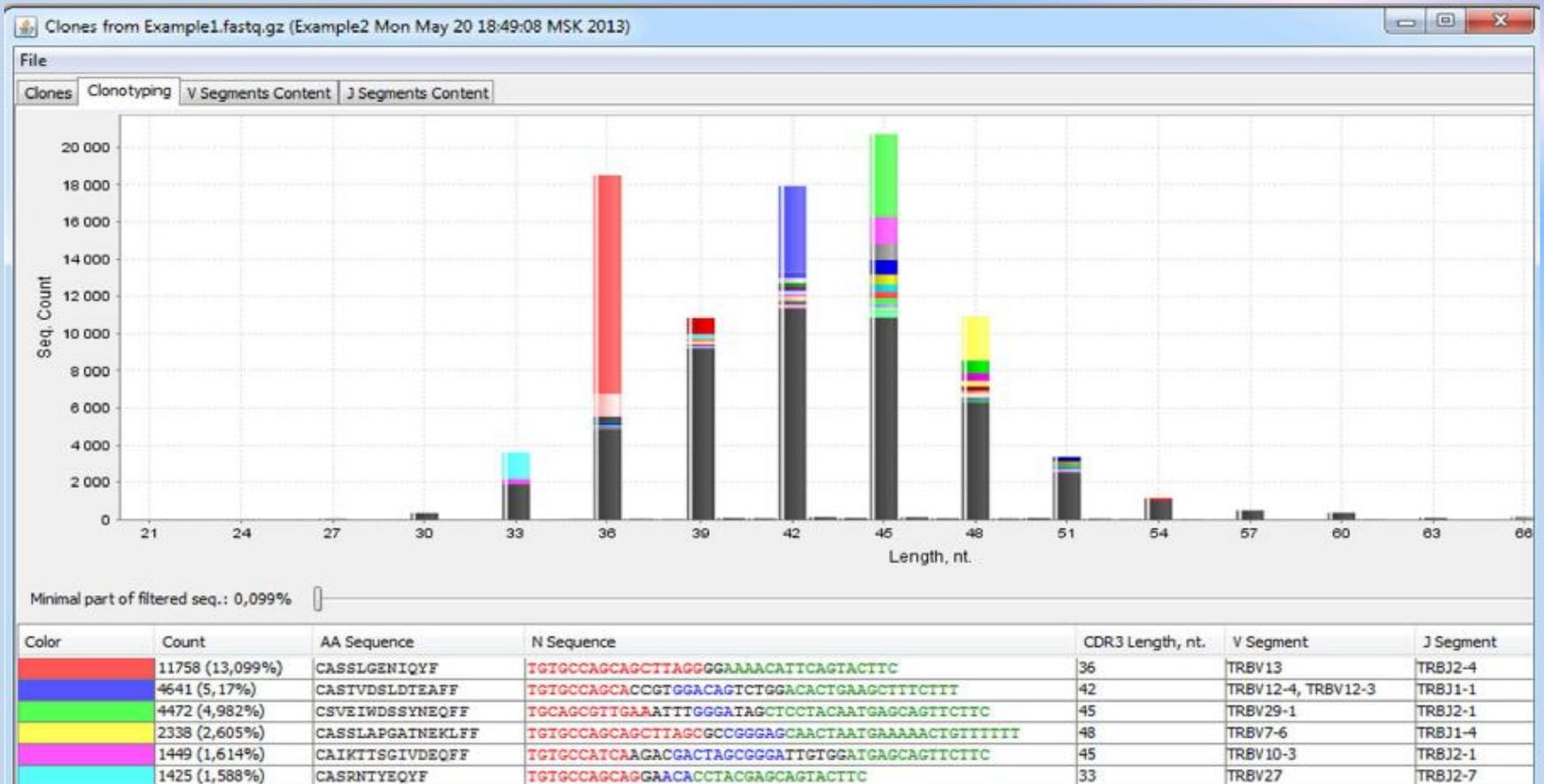
	ASSIGN	ASSIGN	GENDX	GENDX	OMIXON	OMIXON	HLAssign	HLAssign		
SAMPLE	A1	A1	A1	A1	A1	A1	A1	A1	A_SBT	A_SBT
P540	01:01:01:01	24:02:01:01	01:01:01:01	24:02:01:01	01:01:01:01	24:02:01:01	A*01:01:01	A*24:02:01	01:01	24:02
P1138	30:01:01	32:01:01	30:55	32:01:18	30:01:01	32:01:01	A*30:01:01	A*32:01:01	30:01	32:01
P1200	32:01:01	32:01:01	11:178	32:01:01	32:01:01	32:01:01	A*32:01:01	A*32:01:01	32:01	
P1042	01:01:01:01	11:01:01	01:01:01:01	11:01:01	01:01:01:01	11:124/11:39	A*01:01:01	A*11:01:01	01:01	11:01
P1581	02:01:01:01	30:02:01:01	02:01:01:01	30:02:01:01	02:01:01:01	30:02:01	A*02:01:01	A*30:02:01	02:01	30:02
P1728	02:05:01	32:01:01	02:05:01	32:01:01	02:05:01	32:01:01	A*02:05:01	A*32:01:01	02:05	32:01
P964	24:02:01:01	29:02:01:01	24:02:01:01	29:02:01:01	24:02:01:01	29:02:01:01	A*24:02:01	A*29:02:01	24:02	29:02
P959	02:01:01:01	02:05:01	02:05:01	02:81	02:01:01:03	02:02	A*02:01:01	A*02:05:01	02:01	02:05
P978	02:05:01	02:05:01	02:05:01	02:81	02:05:01	02:05:01	A*02:05:01	A*02:05:01	02:05	

	ASSIGN	ASSIGN	GENDX	GENDX	OMIXON	OMIXON	HLAssign	HLAssign		
SAMPLE	DRB1_1	DRB1_1	DRB1_1	DRB1_1	DRB1_1	DRB1_1	DRB1_1	DRB1_1	DRB1_SBT	DRB1_SBT
P540	11:01:01	11:03	11:01:01	11:03	11:01:01	11:03	DRB1*11:01:01	DRB1*11:03	11:01	11:03
P1138	ND	ND	07:01:01:01	16:01:01	16:02:01	16:02:01	DRB1*07:01:01	DRB1*16:01:01	07:01	16:01
P1200	15:01:01 low cov	15:01:01	15:01:01:01	15:01:01:01	07:01:01:01	15:01:01:02	DRB1*07:01:01	DRB1*15:01:01	07:01	15:01
P1042	ND	ND	04:05:01	15:01:01:01	04:05:01	15:02:01:02	DRB1*04:05:01	DRB1*15:01:01	04:05	15:01
P1581	ND	ND	nd	nd	03:01:01:01	03:01:01:02	DRB1*03:01:01	DRB1*03:01:01	03:01	
P1728	ND	ND	11:04:01	16:01:01	11:04:01	16:01:01	DRB1*11:04:01	DRB1*16:01:01	11:04	16:01
P964	ND	ND	01:02:01	10:01:01	01:02:01	10:01:01	DRB1*01:02:01	DRB1*10:01:01	n/d	n/d
P959	16:01:01	16:01:01	16:01:01	16:01:01	16:01:01	16:02:01	DRB1*16:01:01	DRB1*16:01:01	16:01	
P978	ND	ND	nd	nd	16:01:01	16:02:01+04:03:01	no call	no call	04:05	16:01

Risultati

locus	n campioni	concordanza	causa di errore	n alleli osservati
A	48	100%		21
B	96	98,96% o 100%	ambiguità di fase	41
C	96	99,48%	1 errore (SBT?)	24
DRB1	47	100%	1 fallimento	20
DPB1	96	100%		19
DQB1	96	99,48%	esclusione allelica	18
G	40	?		8

TCR e BCR Repertoires



"Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling". Mamedov IZ - *Frontiers in Immunology*, 2013

"Towards error-free profiling of immune repertoires". Shugay M - *Nat Methods*, 2014

Grazie per l'attenzione