BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 4 2009, pages 458–464
doi:10.1093/bioinformatics/btp010

Sequence analysis

Profiling model T-cell metagenomes with short reads

René L. Warren[1,*], Brad H. Nelson[2] and Robert A. Holt[1]

[1]BC Cancer Agency, Michael Smith Genome Sciences Centre, 675 West 10th Avenue, Vancouver, BC V5Z 1L3 Canada and [2]BC Cancer Agency, Deeley Research Centre, 2410 Lee Ave, Victoria, BC V8R 6V5 Canada

Received on September 26, 2008; revised on November 28, 2008; accepted on January 1, 2009

Advance Access publication January 9, 2009

Associate Editor: Alfonso Valencia

Profilování modelových T-buněčných metagenomů s krátkými ready

IV105 - Seminář z bioinformatiky

podzim 2014

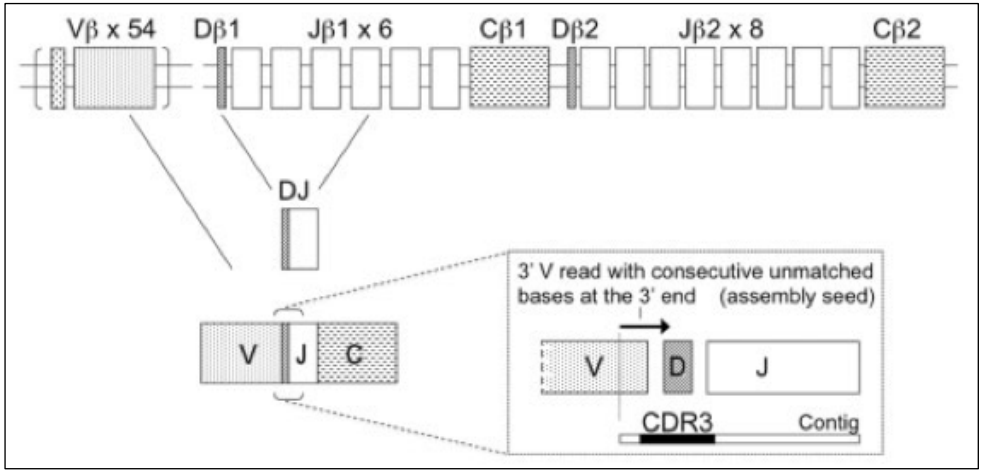
Tomáš Reigl, 357888@mail.muni.cz

Obsah

- TCR (T-cell receptor)
- iSSAKE
- Proces skládání
- Data
- Ukázka

TCR (T-cell receptor)

- vysoká variabilita – až 10^{18} klonotypů (alespoň 10^{7})
- V–(D)–J = CDR3 oblast



Sanger

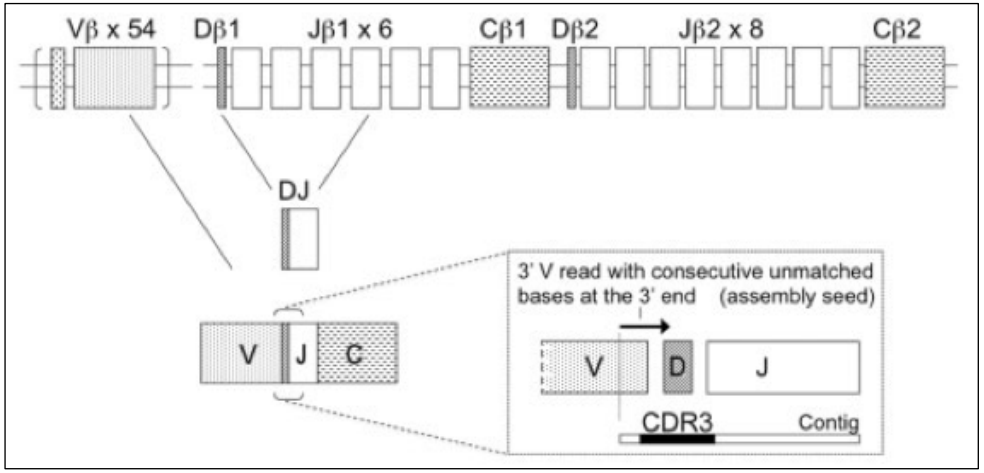
+ dlouhé ready
+ nižší chybovost
− porovnávání délek readů
− vysoká cena
− nízké rozlišení

NGS (Illumina GAII Analyzer)

+ nízká cena
+ vysoké rozlišení
+ konkrétní sekvence
− vyšší chybovost
− krátké ready (nutná rekonstrukce)

TCR (T-cell receptor)

- vysoká variabilita – až 10^{18} klonotypů (alespoň 10^{7})
- V–(D)–J = CDR3 oblast



Sanger

\+ dlouhé ready
\+ nižší chybovost
− porovnávání délek readů
− vysoká cena
− nízké rozlišení

NGS (Illumina GAII Analyzer)

\+ nízká cena
\+ vysoké rozlišení
\+ konkrétní sekvence
− vyšší chybovost
− krátké ready (nutná rekonstrukce)

iSSAKE (immuno-SSAKE)

(SSAKE = Short Sequence Assembly by progressive K-mer search and 30 read Extension (Warren *et al*., 2007))

- nový sekvenační přístup a software
- kompilace krátkých readů z paralelního sekvenování
- profilování metagenomů T-buněk

Vstupní parametry

- -f Fasta file containing all the [paired (-p 1) / unpaired (-p 0)] reads (required)
- ! paired reads must now be separated by ":"
- -s Fasta file containing sequences to use as seeds exclusively (specify only if different from read set, optional)
- -m Minimum number of overlapping bases with the seed/contig during overhang consensus build up (default -m 15)
- -o Minimum number of reads needed to call a base during an extension (default -o 2)
- -r Minimum base ratio used to accept a overhang consensus base (default -r 0.7)
- -t Trim up to -t base(s) on the contig end when all possibilities have been exhausted for an extension (default -t 0)>
- -c Track base coverage for each contig (optional)
- -b Base name for your output files (optional)
- -z Minimum contig size to track base coverage and read position (default -z 50, optional)
- -p Paired-end reads used? (-p 1=yes, -p 0=no, default -p 0)
- -v Runs in verbose mode (-v 1=yes, -v 0=no, default -v 0, optional)
- ========== Options below only considered with -p 1 ==========
- -d Mean distance expected/observed between paired-end reads (default -d 200, optional)
- -e Error (%) allowed on mean distance e.g. -e 0.75 == distance +/- 75% (default -e 0.75, optional)
- -k Minimum number of links (read pairs) to compute scaffold (default -k 2, optional)
- -a Maximum link ratio between two best contig pairs *higher values lead to least accurate scaffolding* (default -a 0.7, optional)
- -g Fasta file containing unpaired sequence reads (optional)

Skládání sekvencí

- vysoký překryv v oblasti V, J a C
- vysoký počet jedinečných sekvencí v CDR3

- využívání anotovaných oblastí V (3’) a J (5’) pro začátky skládání

- přidání náhodné báze do V–D–J spojení
- delece báze na 3’ V
- delece báze na 5’ J

f deleted 3′ V bases			*f* deleted 5′ J bases			*f* added CDR3 bases[a]		
Bases	Observed ($N = 356$)	Simulated ($N = 220\,000$)	Bases	Observed ($N = 1151$)	Simulated ($N = 220\,000$)	Bases	Observed ($N = 174$)	Simulated ($N = 220\,000$)
0	0.194	0.200	0	0.209	0.212	1	0.006	0.007
1	0.160	0.158	1	0.123	0.123	3	0.006	0.006
2	0.098	0.098	2	0.122	0.122	4	0.029	0.031
3	0.118	0.113	3	0.104	0.105	5	0.017	0.019
4	0.160	0.155	4	0.117	0.117	6	0.017	0.019
5	0.118	0.119	5	0.123	0.119	7	0.052	0.056
6	0.070	0.073	6	0.086	0.085	8	0.063	0.067
7	0.045	0.047	7	0.056	0.057	9	0.069	0.073
8	0.022	0.023	8	0.031	0.031	10	0.080	0.084
9	0.008	0.009	9	0.019	0.019	11	0.057	0.059
10	0.006	0.006	10	0.010	0.010	12	0.075	0.076
						13	0.080	0.081
						14	0.086	0.085
						15	0.098	0.096
						16	0.046	0.046
						17	0.052	0.049
						18	0.029	0.028

Testovací soubor

- Genbank TCRβ mRNA
- N = 1 000 000
- 36, 42 nebo 50 nukleotidů dlouhé ready

Clonotype frequency	p.p.m.	Number of unique sTCRβ	Number of total sTCRβ	Fold coverage ca.[a]
1:1 000 000	1	110 000	110 000	5
1:500 000	2	10 000	20 000	10
1:333 333	3	10 000	30 000	15
1:250 000	4	10 000	40 000	20
1:200 000	5	10 000	50 000	25
1:166 667	6	10 000	60 000	30
1:142 857	7	10 000	70 000	35
1:125 000	8	10 000	80 000	40
1:111 111	9	10 000	90 000	45
1:100 000	10	10 000	100 000	50
1:66 667	15	10 000	150 000	75
1:50 000	20	10 000	200 000	100

praktická ukázka

Děkuji za pozornost.