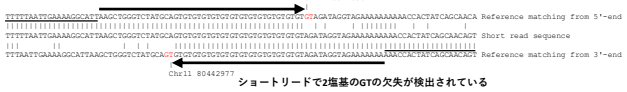
Polymorphic Edge Detection 効率的なNGS配列からの多型検出法

○宮尾安藝雄（農研機構）、清宮健愉（農研機構）、飯田恵子（農研機構）、土井考爾（つくば遺伝子研究所）、安江博（つくば遺伝子研究所）

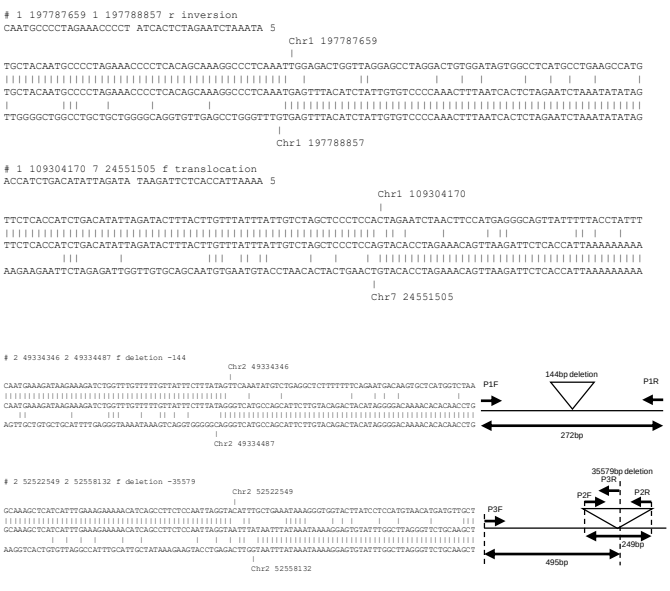
Polymorphic Edgeとは → 多型のキワを検出する

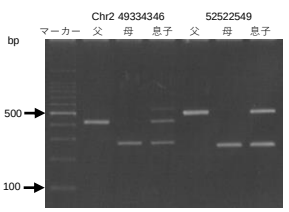
双方向アラインメント法(Bidirectional alignment method)

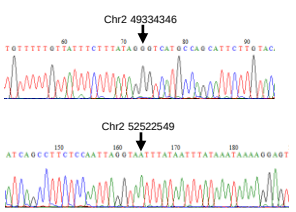


ショートリードで2塩基のGTの欠失が検出されている

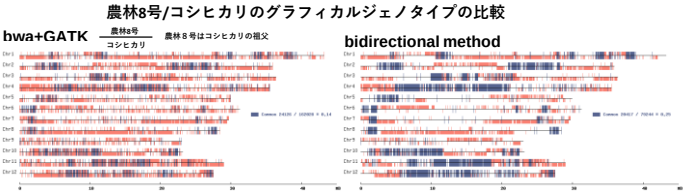
双方向アラインメント法での検出例





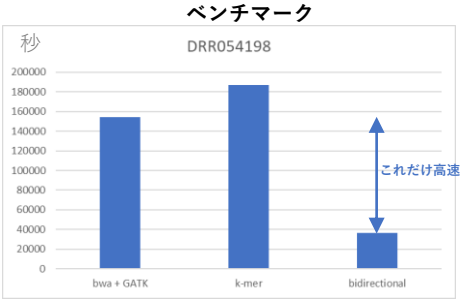


農林8号/コシヒカリのグラフィカルジェノタイプの比較



コシヒカリの農林8号由来のSNP(青)は全SNPに対してGATKでは14%であったが、PEDでは25%で理論値と一致した。

ベンチマーク



一般に用いられているbwa+GATKに比べてbidirectional法は2～2.5倍程度高速に処理できる。K-mer法は非常に大量のk-merの比較処理が必要なので、GATKよりは時間がかかる。

k-merによるキワの検出

k = 20

個体A AAATGGTACATTTATATTAT

個体B AAATGGTACATTTATATTAC

	Name	Value
個体A	AAATGGTACATTTATATTA	T
個体B	AAATGGTACATTTATATTA	C

例
個体Aでは
AAATGGTACATTTATATTATのみ検出され個体Bでは、
AAATGGTACATTTATATTACのみ検出された場合、最後の塩基は多型(SNP、および、構造変異)のエッジである。

k-mer法

a *k* = 20

```
GGTTCCTACTTCAGGGTCATAAAGCCTAAATAGCCCACAC…
GGTTCCTACTTCAGGGTCAT
 GTTCCTACTTCAGGGTCATA
  TTCCTACTTCAGGGTCATAA
   TCCTACTTCAGGGTCATAAA
    CCTACTTCAGGGTCATAAAG
     CTACTTCAGGGTCATAAAGC
      TACTTCAGGGTCATAAAGCC
       ACTTCAGGGTCATAAAGCCT
        CTTCAGGGTCATAAAGCCTA
         TTCAGGGTCATAAAGCCTAA
```

b

Sorted 20-mer	Count
AAAAGATCTATGAGAAGTAG	1
AAAAGATCTATGAGACAGAC	1
AAAAGATCTATGAGACAGAT	50
AAAAGATCTATGAGACATTT	1
AAAAGATCTATGAGATGCAA	89
AAAAGATCTATGAGATGTTG	1
AAAAGATCTATGAGCAAAAC	49
AAAAGATCTATGAGCACTCA	46
AAAAGATCTATGAGCCATCA	1
AAAAGATCTATGAGTAATAT	1

c

19-mer	Last Base A	C	G	T
AAAAGATCTATGAGAAGTA	0	0	1	0
AAAAGATCTATGAGACAGA	0	1	0	50
AAAAGATCTATGAGACATT	0	0	0	1
AAAAGATCTATGAGATGCA	89	0	0	0
AAAAGATCTATGAGATGTT	0	0	1	0
AAAAGATCTATGAGCAAAA	0	49	0	0
AAAAGATCTATGAGCACTC	46	0	0	0
AAAAGATCTATGAGCCATC	1	0	0	0
AAAAGATCTATGAGTAATA	0	0	0	1

d

19-mer	Control A	C	G	T	Target A	C	G	T	Last Base Cont.	Targ.	Genotype
AAAAGATCTATGAGACAGA	0	1	0	50	0	0	0	50	T	T	Wild-type
AAAAGATCTATGAGATGCA	89	0	0	0	107	1	0	0	A	A	Wild-type
AAAAGATCTATGAGCAAAA	0	49	0	0	0	54	0	0	C	C	Wild-type
AAAAGATCTATGAGCACTC	46	0	0	0	40	0	38	0	A	AG	Heterozygous
AATTCATTTTGTGTGGAGG	61	0	0	0	0	0	50	0	A	G	Homozygous

このように*k*-mer法では、マッピングの前に多型を検出できるので、リファレンスゲノムが存在しない生物での多型検出ができる

PEDの特徴

- 次世代シーケンサーの配列データから塩基置換や挿入・欠失変異等の構造変異を効率よく検出します。
- リファレンス配列がない生物の2つのショートリード配列を直接比較してSNPが検出できます(k-mer法)。
- 同一個体の組織間の多型が高精度に検出できます。
- 数100kbを超える大きな欠失も正確に検出します(bidirectional法)。
- 大きなゲノム(コムギ、ヒトの5倍)も解析できます。
- 従来法(bwa+GATK)に比べて、2～3倍高速です(bidirectional法)。
- Linuxが走る普通のコンピュータで解析可能です。
- Fasta配列データに対して単一のスクリプトを実行するだけで、SNPおよび構造変異の検出結果が得られます。

K-mer法による同一人物の正常組織と癌組織でのSNPの検出

Chromosome	Position	(k-1)-mer (k=20)	Reference	SRR2096532	SRR2096535	Direction	Last-base-count SRR2096532(blood) A	C	G	T	SRR2096535(cancer) A	C	G	T	Verify by counting supported reads SRR2096532 Ref	Alt	SRR2096535 Ref	Alt	Genotype
1	17349814	CCTGCCTCTCCAGCACCAT	A	A	AC	r	0	1	1	30	0	1	13	17	11	0	6	4	H
1	29218111	GTAAGTCCCTCTGCTCATG	G	G	AG	f	0	0	29	0	10	0	18	0	35	0	23	10	H
1	29218111	TGGGTCTTGGACACCATGG	G	G	AG	r	0	28	0	0	0	20	0	11	35	0	23	10	H
1	55801587	ATCAGTACAATTTCTCCAC	G	G	CG	r	0	20	0	0	0	16	10	0	22	0	17	10	H
1	57340521	CTGTATTGATTTGTGCAGC	C	C	CT	f	0	27	0	0	0	21	0	10	31	0	19	9	H
1	59092356	GTTATCATGACATCTGAGT	T	T	GT	f	1	0	0	32	0	0	10	21	29	0	19	10	H
1	59243427	TCTTCTGTAAGACCAAAAT	T	C	AC	r	0	0	31	0	0	0	18	12	36	0	25	17	H
1	59243427	TTCTATACTATCTCTGAAA	C	C	AC	f	0	31	0	0	14	24	0	1	36	0	25	17	H
1	69085463	TTTAACCAACAGTGGAAAA	A	A	AG	r	0	0	0	50	0	12	0	45	25	0	19	11	H
1	69350262	TAGAGGAAAATAGCAAAGG	A	A	AG	r	0	0	0	25	0	16	0	26	29	0	29	14	H
1	71920419	CTATGGCTATCACTTCATG	C	C	CT	f	0	27	0	0	0	21	0	15	31	0	19	17	H
1	71920419	CTTGTAATTAAACCTCAAG	C	C	CT	r	0	0	28	0	12	0	21	0	31	0	19	17	H

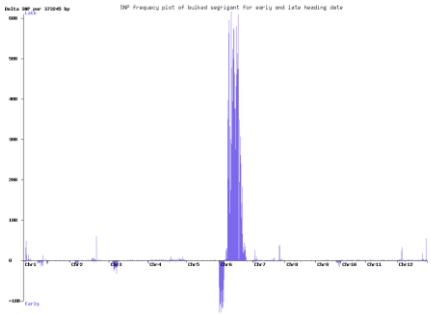
数字は検出されたリード数を示している。

Bidirectional法によるCRISPR/Cas9処理したマウスでの欠失の検出

Accession	Junction by 5' to 3' alignment Chromosome	Position	Junction by 3' to 5' alignment Chromosome	Position	Type	Size	GRCm38 Ref	Alt	Target Ref	Alt	Genotype	Sequence between junction
ERS374107	16	45,512,593	16	45,518,970	deletion	-6,392	10	1	5	3	H	TCAAAACACTTCTT
ERS374108	16	47,502,608	16	47,502,713	deletion	-106	16	1	6	3	H	
ERS374109	17	5,470,888	17	5,471,035	deletion	-152	35	0	10	10	H	CAGA
ERS374108	18	49,793,476	18	49,793,609	deletion	-165	14	0	6	3	H	AGGGAGAGAGGTAAGGAAGGAGGGAGG
ERS374105	19	11,966,045	19	11,966,089	deletion	-56	11	0	5	5	H	GGCGGCGGCGG
ERS374108	19	50,486,921	19	50,529,041	deletion	-42,123	43	0	12	16	H	GT
ERS374109	19	50,486,921	19	50,529,041	deletion	-42,123	43	0	9	14	H	GT
ERS374108	X	89,414,716	X	89,414,796	deletion	-108	15	0	5	3	H	TGTGTGTGTGTGTGTGTGTGTGTGT
ERS374104	X	98,150,098	X	98,150,168	deletion	-58	35	0	11	7	H	TGGGGCGGCCACC
ERS374105	X	98,150,099	X	98,150,174	deletion	-78	35	0	8	6	H	CC
ERS374109	X	98,150,103	X	98,150,171	deletion	-56	35	0	13	8	H	CTGCCTCCCCGAG
ERS374107	X	98,150,120	X	98,150,177	deletion	-60	34	0	12	9	H	GA
ERS374108	X	98,150,120	X	98,150,177	deletion	-60	34	0	10	8	H	GA
ERS374107	X	123,000,254	X	123,662,411	deletion	-662,172	8	1	5	4	H	TTTTTTTTTTTTTT
ERS374106	X	127,841,379	X	127,841,489	deletion	-95	5	1	5	3	H	AACTTTTGGAAGTTAT

黄色の部分がターゲットのアンドロジェンレセプター遺伝子領域

ジャポニカイネとインディカイネの交配後代の早生、晩稲個体バルクの遺伝子型分布図



日本晴(ジャポニカ)とカサラス(インディカ)の交配後代より、早生になった個体と晩稲になった個体をバルクにして解析したNGS配列が公開されている。早生をコントロール、晩稲をターゲットにしてBidirectional法で解析し、変異型を示した座位数の差分(晩稲－早生)をプロットした。第6染色体の上腕部分が早生、その後に晩稲に強く関与する領域が存在することがわかる。

perl bidirectional.pl late early IRGSP1.0
とスクリプトを実行すれば、プロットのもとになる結果が出力される。

論文はこちら

Polymorphic edge detection (PED): two efficient methods of polymorphism detection from next-generation sequencing data
BMC Bioinformatics. 2019 20(1):362.
https://rdcu.be/bH7e8

プログラムはこちら

https://github.com/akiomiyao/ped

日本語のプロトコール

https://tos.nias.affrc.go.jp/~miyao/ngs

Licensed under a Creative Commons表示4.0国際ライセンス
(c)2019 宮尾安藝雄（農研機構）