(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2014-0040697
(43) 공개일자 2014년04월03일

(51) 국제특허분류(Int. Cl.)
C12Q 1/68 (2006.01)
(21) 출원번호 10-2013-7021562
(22) 출원일자(국제) 2012년01월13일
심사청구일자 없음
(85) 번역문제출일자 2013년08월14일
(86) 국제출원번호 PCT/NL2012/050022
(87) 국제공개번호 WO 2012/096579
국제공개일자 2012년07월19일
(30) 우선권주장
61/432,915 2011년01월14일 미국(US)

(71) 출원인
키진 엔.브이.
네덜란드 엔엘-6708 피더블유 웨이제닝엔 아그로비지니스 파크 90
(72) 발명자
반 아이크, 미카엘, 요세푸스, 테레지아
네덜란드, 에이이 와게닝겐 엔엘-6700, 피.오.박스 216 내
(74) 대리인
이원희

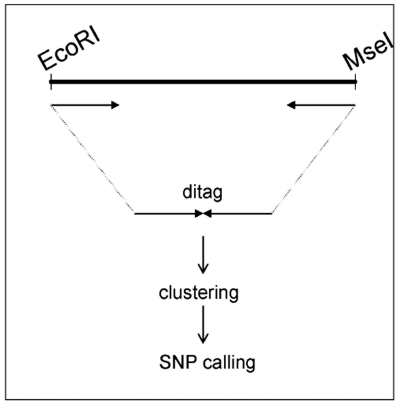
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **쌍 말단 무작위 서열 기반 유전자형 판별**

(57) 요 약

본 발명은 식별자 태그된 제한효소단편들(identifier tagged restriction fragments)을 제공하고, 쌍 고효율 서열분석(high throughput sequencing) 기술들을 이용한 서열 정보를 획득하고, 상기 서열 정보를 결합하고 그리고 샘플들(samples) 간 다형성(polymorphisms)을 확인함으로써 상기 샘플들 간 다형성의 동시 발견, 검출 및 유전자형을 판별(genotyping)하는 방법에 관한 것이다. 양 말단들로부터 서열 정보의 상기 결합은 현저하게 반복적인 유전체들(genomes)에 있는 다형성의 발견, 검출 및 유전자형 판별을 가능하게 한다.

대 표 도 - 도1



특허청구의 범위

청구항 1

하기 단계를 포함하는 하나 또는 그 이상의 또는 다수의 샘플들(samlpes)에 있는 하나 또는 그 이상의 다형성(polymorphisms)의 동시 발견, 검출, 및 유전자형 판별(genotyping) 방법;

(a) 하나 또는 그 이상 또는 다수의 샘플들로부터 DNA를 제공하는 단계;

(b) 제한효소단편들(restriction fragments)을 생산하기 위해 적어도 하나의 제한 엔도뉴클레아제(restriction endonuclease)를 가지고 상기 DNA를 절단함으로써 상기 샘플의 복잡도를 줄이는 단계;

(c) 태그된(tagged) 제한효소단편들을 생산하기 위해 적어도 하나의 식별자 태그(tag)를 갖는 샘플의 상기 제한효소단편들을 제공하는 단계;

(d) 상기 태그된 제한효소단편들의 적어도 일부분을 쌍-말단(paired-end) 서열분석(sequencing)하는 단계; 및

(e) 상기 샘플들 간에 다형성을 확인하는 단계.

청구항 2

제 1항에 있어서, 상기 단편의 쌍 말단 서열분석된 리드들(reads)의 제 1 서열 리드 및 제 2 서열 리드는 다이태그(ditag), 바람직하게 인 실리코(in silico)에 결합되는 것을 특징으로 하는 방법.

청구항 3

제 1항 또는 제 2항에 있어서, 상기 제 1 또는 제 2 서열 리드 중 어느 하나는 다이태그에 결합하기 전 역 상보되는(reverse complemented) 것을 특징으로 하는 방법.

청구항 4

제 1항 내지 3항에 있어서, 상기 식별자 태그는 하기에 의해 제공되는 것을 특징으로 하는 방법:

- 태그된 어탭터(adaptor) 라이게이션된(ligated) 제한효소단편들을 생산하기 위해 제한효소단편들에 태그된 어탭터들을 라이게이션하는 것; 또는

- 태그된 어탭터 라이게이션된 제한효소단편들을 생산하기 위해 상기 어탭터의 적어도 일부분과 상보적인 적어도 하나의 태그된 프라이머(primer)를 가지고 어탭터 라이게이션된 제한효소단편들을 증폭하는 것.

청구항 5

제 1항 내지 4항에 있어서, 상기 서열들은 상기 식별자 태그에 기반된 상기 샘플들에 할당되는 것을 특징으로 하는 방법.

청구항 6

제 1항 내지 5항에 있어서, 상기 할당된 서열들은 상기 샘플들 간 서열들에 있는 다형성의 확인을 위해 샘플들을 비교하는 것을 특징으로 하는 방법.

청구항 7

제 1항 내지 6항에 있어서, 상기 다이태그들은 상기 샘플들 간 비교하는 것을 특징으로 하는 방법.

청구항 8

제 1항 내지 7항에 있어서, 상기 샘플들은 확인된 다형성에 근거하여 유전자형을 판별하는 것을 특징으로 하는 방법.

청구항 9

제 1항 내지 8항에 있어서, 상기 복잡도 감소는 제한효소단편들을 생산하기 위해 둘 또는 그 이상의 제한 엔도뉴클레아제를 가지고 상기 샘플 DNA의 절단을 포함하는 것을 특징으로 하는 방법.

청구항 10

제 1항 내지 9항에 있어서, 상기 어탭터들은 어탭터 라이게이션된 단편들을 제공하기 위해 제한효소단편들의 하나 또는 양 말단들에 라이게이션되는 것을 특징으로 하는 방법.

청구항 11

제 9항 또는 10항에 있어서, 다른 제한효소(restriction enzyme)에 의해 획득되는 제한효소단편의 상기 각 말단을 위해, 다른 어탭터가 라이게이션되는 것을 특징으로 하는 방법.

청구항 12

제 10항 또는 11항에 있어서, 상기 복잡도 감소는 추가적으로 상기 어탭터의 일부분과 적어도 상보적인 적어도 하나의 프라이머를 가지고 상기 어탭터 라이게이션된 단편들을 증폭하는 것을 포함하는 방법.

청구항 13

제 12항에 있어서, 상기 프라이머는 추가적으로 상기 제한 엔도뉴클레아제의 인식 서열의 남아 있는 부분의 적어도 일부분과 상보적인 것을 특징으로 하는 방법.

청구항 14

제 13항에 있어서, 상기 프라이머는 추가적으로 프라이머의 3' 말단에 하나 또는 그 이상의 무작위적으로 선택된 뉴클레오티드(nucleotides)를 포함하는 것을 특징으로 하는 방법.

청구항 15

제 13항에 있어서, 상기 프라이머는 하나 또는 그 이상의 샘플들을 위한 프라이머의 3' 말단에 동일한 하나 또는 그 이상의 무작위적으로 선택된 뉴클레오티드를 포함하는 것을 특징으로 하는 방법.

청구항 16

상기 항 중 어느 한 항에 있어서 상기 서열분석은 고효율(high throughput) 서열분석에 기반되는 것을 특징으로

하는 방법.

청구항 17

제 13항에 있어서, 상기 고효율 서열분석은 파이로시퀀싱(pyrosequencing), 바람직하게 고체 담체(solid carrier)에 디(d)에 기반되는 것을 특징으로 하는 방법.

청구항 18

제 13항에 있어서, 상기 고효율 서열분석은 라이게이션, 또는 나노세공(nanopore) 서열분석에 의한 서열분석에 기반되는 것을 특징으로 하는 방법.

명 세 서

배 경 기 술

[0001] 현재 전형적으로 사용되고 있는 주요한 마커(marker) 발견 및 유전자형 판별(genotyping) 기술들은 두 가지 다른 시스템들, SNPs의 초기 발견을 위한 기술 및 그 뒤에 수많은 개인들을 유전자형 판별하기 위한 또 다른 기술에 의존한다. 이것은 무작위 서열 기반 유전자형 판별(random sequence-based genotyping, rSBG)로 불리는, 동시 서열-기반된 마커 발견 및 검출을 위한 기술을 발달시키기 위한 본 출원을 상기시켰다. 상기 기술은 Illumina GA*II*의 고효율(high-throughput) 서열분석 수용량 및 AFLP®(EP534858)의 유전체(genome) 복잡도 감소 수용량을 포함하고 있다. 상기의 예는 본 출원에 의해 WO2007073165에 설명되어 있다. 종종 표적화 되는 다양한 다른 유전자형 판별 기술들(예를 들어., 검출되는 SNPs는 먼저 선택되고 그리고 특정 검출 프로브들(probes)을 이용함으로써 표적화 된다.)과 달리, rSBG는 *무작위* 접근법이다. 이론적으로 라인들(lines) 간 현존하는 모든 SNPs는 그리고 AFLP 주형들에 특정 서열들이 포함되어 있을 때, (전형적으로 엄격한 탐색 필터들(mining filters)을 적용한 후) 분류될 수 있다. 문제들 중 하나는 후추와 같은, 반복적인 서열들의 상대적으로 큰 부분을 포함하는 유전체들로부터 유래한 샘플들을 분석될 때, 라인들 간 다형성(polymorphisms)의 확인이 반복적인 서열들의 존재 때문에 점점 더 어려워진다.

발명의 내용

[0002] 본 발명자들은 다수의 샘플들에서 점수화되고 그리고 유전자형이 판별되는 다형성(polymorphisms)의 수에 있어서 개선들이 달성될 수 있고, 그리고 실직적으로 샘플들이 현저히 반복적인 것으로 고려되는 유전체들(genomes)로부터 사용될 때, 예를 들어 많은 반복 서열을 포함할 때, 고효율 서열분석(high throughput sequencing) 방법들이 제한효소단편(restriction fragments)의 양 말단을 서열분석하기 위해 사용될 때 달성될 수 있음을 밝혔다. 소위 쌍-말단 서열분석 접근법들(paired-end sequencing approaches)이라 불리는 것을 도입함으로써, 두 세트의 서열 데이터(aka 서열 리드들(reads))는 동일한 제한효소단편으로부터 획득되고, 하나는 제한효소단편의 각 말단으로부터 획득된다. 상기 세트들을 결합함으로써, 서로가 달리 구별될 수 없는 제한효소단편으로부터의 서열 데이터는, 예를 들어 상기 서열 데이터는 반복서열들로부터 유래하기 때문에, 이제 구별될 수 있다. 그 이유는 (종종, 사용되는 제한효소들(restriction enzymes) 또는 단편화 방법에 의존하는) 수백 또는 수천 개의 뉴클레오티드(nucleotide)에 떨어져 위치해 있는), 상기 제한효소단편의 다른 말단으로부터의 서열 리드가 결합된 고유의 서열 리드들을 나타낼 수 있기 때문이다(도 1 참조). 이는 이제 또한 현저히 반복적인 유전체들로부터 획득된 샘플들의 다형성의 발견, 검출 및 유전자형 판별을 가능하게 한다. 광범위하게, 본 발명의 방법은 따라서 성공적으로 현저하게 반복적인 샘플들을 포함함으로써 이전 기술의 방법들보다 샘플들을 더 광범위하게 적용할 수 있다. 본 발명자들은 더 많은 SNPs가 단편들의 각 말단의 리드들의 별개의 분석들과 비교되는, 본 발명의 쌍-말단(paired-end) 접근법을 가지고 발견될 수 있고 그리고 유전자형이 판별될 수 있고, 예를 들어, SNPs의 동시 발견 및 유전자형 판별에서 상기 쌍-말단 서열분석의 사용에 기인할 수 있는 상승 효과(synergistic effect)가 획득되는 것을 밝혔다.

도면의 간단한 설명

[0003] 도 1: 쌍-말단(paired-end) 무작위 서열 기반 유전자형 판별(genotyping)의 도식 표현. 다이태그들(Ditags)은 SNP 확인을 위해 반복 서열의 최대 분리를 달성하기 위하여 제한효소단편들(resctriction fragments)의 반복적인 말단으로부터 만들어졌다.

발명을 실시하기 위한 구체적인 내용

[0004] **정의(Definitions)**

[0005] 하기 설명 및 실시예에서, 수많은 정의들이 사용되었다. 명세서 및 하기 정의를 특정하는 범위를 포함한, 청구항들의 명확하고 일관된 이해를 제공하기 위하여, 하기 정의들이 제공된다. 본 명세서에 정의되지 않은, 사용된 모든 기술적인 그리고 과학적 정의들은 당해 기술 분야에 숙련된 사람에 의해 일반적으로 이해될 수 있는 동일한 의미이다. 모든 출판, 특허 출원, 특허 및 다른 참조의 공개는 본 명세서에서 참조에 의해 이들 전부에 포함된다.

[0006] 본 발명의 방법으로 사용된 고식적 기술들을 수행하는 방법은 당해 기술 분야에 숙련된 사람에게 명확해질 수 있다. 분자 생물학, 생화학, 컴퓨터화학(computational chemistry), 세포배양, 재조합 DNA, 생물정보학, 유전체학, 서열분석 및 관련된 분야에서 고식적 기술들의 수행은 당해 기술 분야에 숙련된 사람들에 의해 잘 알려져 있고 그리고 , 예를 들어, 하기 문헌 참조: Sambrook et al. ., Molecular Cloning. A Laboratory Manual, 2nd Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N. Y., 1989; Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, New York, 1987 and periodic updates; and the series Methods in Enzymology, Academic Press, San Diego에서 논의되고 있다.

[0007] 여기에 사용된, 단수 "하나(a),""하나(an)" 및 "그(the)"는 문맥이 명확하게 명시하지 않는 한 복수 지시대상들을 포함한다. 예를 들어, "한 개(a)" DNA 분자"를 분리하기 위한 방법은 다수의 분자들을 분리하는 것을 포함한다(예컨대, 수십, 수백, 수천, 수 십만, 수 백만, 또는 그 이상의 분자들).

[0008] 다형성(Polymorphism): 다형성은 개체 내 뉴클레오티드 서열(nucleotide sequence)의 둘 또는 그 이상의 변종의 존재를 말한다. 다형성은 하나 또는 그 이상의 염기 치환(base change), 삽입, 반복, 또는 결실을 포함한다. 다형성은 예컨대 단순 서열 반복(simple sequence repeat, SSR) 및 단일 뉴클레오티드: 아데닌(adenin, A), 티민(thymine, T), 시토신(cytosine, C), 또는 구아닌(guanine, G)이 바뀌었을 때를 포함한, DNA 서열 변종인, 단일 뉴클레오티드 다형성(single nucleotide polymorphism, SNP)을 포함한다. 변형은 일반적으로 SNP가 고려되는 적어도 1%의 개체군 내에서 나타난다. SNPs는 예컨대 모든 인간 유전자 변형(genetic variactions)의 90%를 이루고, 그리고 인간 유전체(genome)를 따라 모든 100 개 내지 300 개의 염기들에서 나타난다. 모든 세 개의 SNPs 중 둘은 시토신(C)이 티민(T)으로 대체된다. 예컨대 인간들 또는 식물들의 DNA 서열에 있어서 변종들은 어떻게 변종들이 질병들, 박테리아, 바이러스들, 화학물질들, 약물들, 등을 다루는지에 영향을 미칠 수 있다.

[0009] 유전자형 판별(Genotyping)은 종들 내 개인들 간 유전자 변형을 결정하는 과정을 말한다. 생물의 유전자형(genotype)은 유전자 코드(genetic code) 내에 유전자형이 지니는 유전 지시들이다. 모습 및 행동이 환경적 그리고 발달된 상태에 의해 변형되므로 동일한 유전자형을 가진 모든 생물들이 동일한 방법으로 보이거나 또는 행동하는 것은 아니다. 마찬가지로, 필수불가결하게 비슷하게 보이는 모든 생물들이 동일한 유전자형을 가지는 것은 아니다. 단일 뉴클레오티드 다형성(SNPs)은 가장 일반적인 유전자 변형의 형태이고 그리고 정의에 의해 1% 이상의 개체에서 발견되는 특정 유전자좌(locus)에서의 단일-염기 차이들이다. SNPs는 유전체의 코딩(coding) 및 비코딩(non-conding) 지역 모두에서 발견되고 그리고 코딩 지역에서 발견되었을 때, 질병을 유발하는 또는 질병에 저항력을 가지는 능력과 같은, 다른 유전자형을 유도할 수 있다. 그러므로, SNPs는 종종 특정 질병들 또는 몇몇 유전자형들을 위한 마커들(markers)로 사용된다. 비코딩 지역에서 발견되었을 때, SNPs는 진화 유전체학 연구를 위한 마커들로서 역할을 한다. "InDels" 또는 다양한 길이의 뉴클레오티드의 삽입 및 결실은 SNPs와 관련이 있다. 유전자 변형의 세 번째 형태는 복제수 변이(copy number variation, CNV)로, 이것은 다양한 유전체 내에 DNA 분절(segment)의 다른 복제 수를 가짐으로부터 야기한다. 상기 복제수 변형이 코드되는 유전자에서인 경우, 상기 변형은 질병에 대한 감수성 또는 저항성을 유도할 수 있다. 몇몇 유전자형들은 또한 정량 민감성(dosage-sensative)이고, 그리고 상기 복제수는 종의 멤버들(members) 간 가변성을 상기시키는 원인이다. SNP 및 CNV 유전자형 판별 모두를 위하여, 많은 방법들이 개인들 간 유전자형을 결정하기 위해 존재한다. 상기 선택된 방법은 일반적으로 유전자형이 판별되는 개인들의 수 및 각 개인을 위해 검정되는 유전자형

의 수 모두의 기능인, 처리율 요구(throughput needs)에 의존한다. 상기 선택된 방법은 또한 각 개인 또는 샘플로부터 이용 가능한 샘플 재료의 양에 의존한다.

[0010] 상기 유전자형은 보통 특정 특징 또는 고려 중인 특성에 관련하여 세포, 생물, 개인의 유전자 구성(예를 들어 개인의 특정 대립유전자 구성)이다.

[0011] 유전자형은 형태학, 발달, 생화학적 또는 생리학적 특징들, 생물 계절학(phenology), 행동, 및 행동의 산물들과 같은 생물의 관찰 가능한 특징들 또는 특성들이다. 유전자형들은 둘 간 환경적인 요인들 및 상호작용들의 영향뿐만 아니라 유전자들의 발현으로부터 유래한다. 비록 유전자형이 생물에 의해 드러나는 관찰 가능한 특징들의 총체이지만, 단어 페놈(phenome)이 종종 특성들의 집합을 말하기 위해 사용되고 그리고 상기 특성들의 동시 연구는 페노믹스(phenomics)라 불린다.

[0012] 유전자형 판별은 생물의 유전자형을 결정하는 것이다.

[0013] 제한 엔도뉴클레아제(Restrection endonuclease): 제한 엔도뉴클레아제 또는 제한효소(restriction enzyme)는 이중가닥 DNA 분자에 있는 특정 뉴클레오티드 서열(표적 부위(target site))를 인식하는 효소이며, 그리고 평활 또는 접착성 말단을 남기면서, 모든 표적 부위에서 또는 근처에서 DNA 분자의 양 가닥을 절단할 수 있다.

[0014] 제한효소단편들(Restrection fragments): 제한 엔도뉴클레아제를 가지고 절단에 의해 생산된 상기 DNA 분자들은 제한효소단편들이라 불린다. 어느 주어진 유전체(또는 핵산, 기원에 상관 없는)도 특정 제한 엔도뉴클레아제에 의해 제한효소단편들의 별개의 세트로 절단될 수 있다. 제한 엔도뉴클레아제 절단으로부터 야기한 상기 DNA 단편들은 추가적으로 다양한 기술들에 사용될 수 있다.

[0015] 태깅(Tagging): 여기에 정의된 태깅이라는 용어는 두 번째 또는 추가 핵산 샘플로부터 핵산 샘플을 구별할 수 있도록 하기 위하여 상기 핵산 샘플에 태그(tag)를 추가하는 것을 말한다.

[0016] 식별자(Identifer) 또는 식별자 태그(identifier tag): 어탭터(adaptor) 또는 프라이머(primer)에 추가될 수 있는 또는 서열에 포함될 수 있는 또는 고유의 식별자를 제공하기 위한 라벨(label)로 사용될 수 있는 짧은 서열. 이런 서열 식별자(태그)는 다양한 그러나 전형적으로 4 내지 16 bp로, 정의된 길이의 고유의 염기 서열일 수 있다. 상기 식별자, 또는 식별자들의 조합들은 특정 핵산 샘플을 확인하기 위하여 또는 DNA 산물, 예를 들어 샘플에서 유래한 단편 또는 PCR-산물을 연결하거나 또는 연관짓기 위해 사용된다. 예를 들어 4 bp 태그는 4승(exp4) = 256개 다른 태그들을 가능하게 한다. 이런 식별자를 사용하여, 샘플의 기원이 추가의 처리 중에 결정될 수 있다. 다른 핵산 샘플들로부터 유래한 처리된 산물들을 결합하는 경우, 상기 다른 핵산 샘플들은 일반적으로 다른 식별자들을 이용하여 확인된다. 식별자들은 바람직하게 적어도 두 개의 염기 쌍에 의해 각각이 서로 다르고 그리고 바람직하게 잘못된 리드들(misreads)을 막기 위해 두 개의 동일한 연속적인 염기들을 포함하지 않는다. 상기 식별자의 기능은 종종 어탭터들 또는 프라이머들과 같은 다른 기능들과 결합할 수 있다.

[0017] 태그된 제한효소단편(Tagged restrection fragment): 식별자 태그와 함께 제공되는 제한효소단편.

[0018] 어탭터 라이게이션된 제한효소단편들(Adaptor-ligated restriction fragments): 어탭터들에 의해 감싸진 제한효소단편들.

[0019] 어탭터들(Adaptors): 제한된 수의 염기쌍, 예컨대 길이에 있어서 대략 10 내지 대략 30 염기쌍을 가지는 짧은 이중가닥 DNA 분자들로, 제한효소단편들의 말단들에 라이게이션될 수 있도록 고안되었다. 어탭터들은 일반적으로 서로 일부분 상보적인 뉴클레오티드 서열들을 갖는 두 개의 합성 올리고뉴클레오티드(oligonucleotides)로 구성되어 있다. 적절한 상태 하에 용액 내에 상기 두 개의 합성 올리고뉴클레오티드들을 혼합할 때, 상기 올리고뉴클레오티드들은 서로 어닐(anneal)하여 이중가닥 구조를 형성할 수 있다. 어닐링(annealing) 후, 상기 어탭터 분자의 한쪽 말단은 제한효소단편의 말단과 경쟁할 수 있도록 고안되고 그리고 상기와 같이 라이게이션될 수 있다; 상기 어탭터의 다른 쪽 말단은 라이게이션될 수 없도록 고안될 수 있지만, 경우(이중 라이게이션된 어탭터들)에 따라 필요하지 않다.

[0020] 라이게이션(Ligation): 두 개의 이중가닥 DNA 분자들이 서로 공유결합으로 연결되도록 하는 리가아제(ligase) 효소에 의해 촉매되는 효소 반응을 라이게이션이라고 말한다. 일반적으로, 두 개의 DNA 가닥 모두 서로 공유결합으로 결합되지만, 또한 상기 가닥들의 말단들 중 하나의 화학적 또는 효소적 변형을 통해 두 개의 가닥들 중 하나의 라이게이션을 막는 것이 가능하다. 상기의 경우 공유결합 연결은 상기 두 개의 DNA 가닥들의 오직 하나에서 일어날 수 있다.

[0021] 프라이머들(Primers): 일반적으로, 여기에 정의된 프라이머들이라는 용어는 DNA의 합성을 준비할 수 있는 DNA

가닥들을 말한다. DNA 중합효소(polymerase)는 프라이머 없이 처음부터 DNA를 합성할 수 없다: DNA 중합효소는 오직 상보적인 가닥이 조립되는 뉴클레오티드들의 순서를 지시하기 위한 주형으로서 사용되는 반응에서 존재하는 DNA 가닥을 연장할 수 있다. 우리는 중합효소 연쇄 반응(polymerase chain reaction, PCR)에 사용되는 합성 올리고뉴클레오티드들을 프라이머들이라고 말할 수 있다.

[0022] 합성 올리고뉴클레오티드(Synthetic oligonucleotide): 화학적으로 합성될 수 있는, 바람직하게 대략 10 내지 대략 50 염기들을 가지는 단일가닥 DNA 분자들은 합성 올리고뉴클레오티드라고 불린다. 일반적으로, 비록 관련된 서열들을 가지는 그리고 뉴클레오티드 서열 내에 특정 위치에서 다른 뉴클레오티드 구성을 가지는 분자들의 계통을 합성할 수 있지만, 상기 합성 DNA 분자들은 고유의 또는 바라는 뉴클레오티드 서열을 가지도록 고안된다. 여기에 정의된 합성 뉴클레오티드라는 용어는 고안된 또는 바라는 뉴클레오티드 서열을 가지는 DNA 분자들을 말하는데 사용될 수 있다.

[0023] 증폭(Amplification): 여기에 정의된 증폭이라는 용어는 전형적으로 PCR을 사용하여, 이중가닥 DNA 분자들의 생체 외(in vitro) 합성을 나타내는데 전형적으로 사용될 수 있다. 다른 증폭 방법들이 존재하고 그리고 상기 방법들이 요지에 벗어나지 않고 본 발명에서 사용될 수 있음이 잘 알려져 있다.

[0024] 앰플리콘(Amplicon): 폴리뉴클레오티드(polynucleotide) 증폭 반응의 산물, 즉, 하나 또는 그 이상의 시작 서열로부터 복제되는 폴리뉴클레오티드 집단. 앰플리콘은 중합효소 연쇄 반응들(PCRs), 선형 중합효소 반응들(linear polymerase reactions), 핵산 서열 기반 증폭, 회전환 증폭(rolling circle amplification) 및 이와 같은 반응들을 포함하지만 이에 한정되지 않는, 다양한 증폭 반응에 의해 생산될 수 있다.

[0025] 복잡도 감소(Complexity reduction): 여기에 정의된 복잡도 감소라는 용어는 유전체 DNA와 같은, 핵산 샘플의 상기 복잡도가 샘플의 부분집합(subset)의 발생에 의해 감소 되는 방법을 말하기 위해 사용된다. 상기 부분집합은 전체(예를 들어 복합체(complex)) 샘플을 대표할 수 있고 그리고 바람직하게 복사할 수 있는(reproducible) 부분집합이다. 복사할 수 있는(Reproducible)은 본 문단에서 동일한 샘플이 동일한 방법을 이용하여 복잡도에 있어서 감소될 때, 동일한 또는 적어도 비슷한, 부분집합이 얻어지는 것을 의미한다. 복잡도 감소를 위해 사용되는 상기 방법은 당해 기술 분야에서 알려진 복잡도 감소를 위한 방법 중 어느 하나일 수 있다. 복잡도 감소를 위한 방법들의 예들은 예를 들어 AFLP®(Keygene N.V., the Netherlands; 예컨대 EP 0 534 858 참조), Dong(예컨대 WO 03/012118, WO 00/24939 참조)에 설명된 방법들, 색인용 링킹(indexed linking)(Unrau, et al., 1994, Gene, 145:163-169), 등을 포함한다. 본 발명에 사용된 상기 복잡도 감소 방법들은 복사할 수 있다는 공통점을 지닌다. 현미 해부(microdissction) 또는 선택된 조직에서 그리고 복제를 위해 전사되는 유전체의 일부를 나타내는 mRNA(cDNA)의 사용과 같은 더 무작위적 복잡도 감소와 대조적으로, 동일한 샘플이 동일한 방식으로 복잡도에 있어서 감소될 때, 동일한 샘플의 동일한 부분집합이 얻어진다는 점에서 복사할 수 있는(reproducible)은 조직의 선택, 분리 시간 등에 의존한다.

[0026] 선택적인 염기(Selective base), 선택적인 뉴클레오티드(selective nucleotide), 무작위로 선택적인 뉴클레오티드(randomly selective nucleotide): 프라이머의 3' 말단에 위치된, 선택적인 염기는 무작위로 A, C, T 또는 G(또는 경우에 따라 U) 사이에서 선택된다. 선택적인 염기를 가지고 프라이머를 연장함으로써, 차후 증폭은 어탭터 라이게이션된 제한효소단편들, 예를 들어 선택적인 염기를 전달하는 프라이머를 사용하여 증폭될 수 있는 유일한 단편들의 유일하게 복사할 수 있는 부분집합을 산출할 수 있다. 선택적인 뉴클레오티드는 1과 10 사이의 다양한 수로 프라이머의 3' 말단에 추가될 수 있다. 전형적으로 1 내지 4가 충분하다. (PCR에서) 프라이머들 둘 다 다양한 수의 선택적인 염기들을 포함할 수 있다. 각각의 추가된 선택적인 염기들과 함께, 상기 부분집합은 부분집합에서 증폭된 어탭터-라이게이션된 제한효소단편들의 양을 대략 4의 값으로 감소시킨다. 복잡도 감소의 상기 형태는 이전 서열 지식 중 어느 하나도 요구하거나 또는 고려하지 않고, 오직 선택적인 뉴클레오티드에 기반함으로써 무작위한 것으로 고려된다. 전형적으로, AFLP 기술(EP534858)에 사용된 선택적인 염기들의 수는 +N+M에 의해 표시되며, 상기 하나의 프라이머는 N 선택적인 뉴클레오티드들을 전달하고 그리고 다른 하나의 프라이머들은 M 선택적인 뉴클레오티드들을 전달한다. 그러므로, Eco/Mse +1/+2 AFLP는 EcoRI 및 MseI을 가지고 시작 DNA를 절단, 적절한 어탭터들의 라이게이션 및 하나의 선택적인 염기를 전달하는 EcoRI 제한적인 위치에서 있는 하나의 프라이머 및 2개의 선택적인 뉴클레오티드를 전달하는 MseI 제한적인 위치에 있는 다른 하나의 프라이머를 가지고 증폭하는 것의 약칭이다. 3' 말단에 적어도 하나의 선택적인 뉴클레오티드를 전달하는 AFLP에 사용되는 프라이머는 또한 AFLP-프라이머로서 묘사된다. 3'에서 선택적인 뉴클레오티드를 전달하지 않고 그리고 사실 어탭터 및 제한효소 절단부위의 나머지들과 상보적인 프라이머들은 종종 AFLP+0 프라이머들로 나타내어진다. 여기에 정의된 선택적인 뉴클레오티드라는 용어는 또한 어탭터 부분(section)에 인접해서 위치한 그리고 상기의 결과로서, 뉴클레오티드가 알려지게 되는 선택적인 프라이머의 사용에 의해 확인되는 표적 서

열의 뉴클레오티드들을 위해 사용된다.

[0027] 서열분석(Sequencing): 여기에 정의된 서열분석이라는 용어는 핵산 샘플, 예컨대 DNA 또는 RNA에서 뉴클레오티드들(염기 서열들)의 순서를 결정하는 것을 말한다. 많은 기술들은 생거(Sanger) 서열분석 그리고 둘 다 파이로시퀀싱(pyrosequencing)에 기반된, Roche Applied Science에 의해 제공되는 GS FLX 플랫폼(platform)과 같은 고효율(high-throughput) 서열분석 기술들(또한 차세대 염기서열분석(next-generation sequencing) 기술들로서 잘 알려진), 및 Illumina의 유전체 분석기(Genome Analyzer)와 같이 이용 가능하다. 다른 플랫폼들도 존재한다.

[0028] 고효율 서열분석 또는 차세대 염기서열분석은 많은 양의 리드들(reads), 전형적으로 동시에 몇 백보다 많은 수천(예를 들어 수 십 또는 수 백만) 또는 수많은 서열 리드들의 순서를 발생시킬 수 있는 서열분석 기술이다. 고효율 서열분석은 고식적 생거 또는 모세관 서열분석(capillary sequencing)으로부터 구별되고 그리고 분명하다. 전형적으로, 서열화된 산물들은 전형적으로 대략 600 및 30 bp 사이에서, 상대적으로 짧은 리드들을 가지는 서열화된 산물들이다. 상기 방법들의 예들은 Seo et al. (2004) Proc. Natl. Acad. Sci. USA 101:5488-93에 의해, WO 03/004690,WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, 및 WO 2005/003375에 공개된 파이로시퀀싱-기반된 방법들에 의해 제시된다. 상기 기술들은 전형적으로 추가적으로 광범위하고 그리고 정교한 데이터 저장 및 리드 조립(read assaembly) 등을 위한 처리 작업흐름(workflows)를 구성한다. 고효율 서열분석의 이용가능성은 많은 고식적 작업흐름 및 현재 생산되는 데이터의 형태 및 품질을 수용하기 위해 재고안된 유전체의 분석을 위한 방법을 요구한다.

[0029] 여기에 사용된, '쌍 말단 서열분석(paired end sequencing)'은 Illumina 및 Roche에 의해 현재 팔리는 플랫폼들에 부분적 기반된, 고효율 서열분석에 기반된 방법이다. Illumina는 업그레이드(upgrade)로서 존재하는 시퀀서(sequencer)에 설치될 수 있는 하드웨어 모듈(hardware module)(PE 모듈)을 공개했고, 이는 주형의 양 말단의 서열분석을 가능하게 하고, 따라서 쌍 말단 리드들을 발생시킨다. 쌍 말단 서열분석은 "Next generation sequencing: towards personalised medicine. Michael Janitz Ed., 2008, Wiley section 2.4에서 Lakdawalla에 의해 설명된 바와 같이, 서열분석이 수행되는 캐리어(carrier) 상에서 서열분석하는 DNA 분자의 가닥의 재배향(reorientation)에 의해 완성된다. 상기 쌍 말단 서열분석의 형태는 전형적으로 더 짧은 단편들(대략 1000 bp 이상)을 위해 사용된다. 또다른 쌍 말단 서열분석의 변형은 종종 메이트-페어(mate-pair) 서열분석으로서 나타내어지고, 상기 서열분석 어탭터들은 DNA 단편들에 라이게이션 되며, 라이게이션된 DNAs는 인식 서열이 어탭터에 포함되어 있는 타입 IIs(type IIs) 제한효소들에 의해 절단되고, 자가-환형되고(self-circularised), 타입 IIs가 소화되고 그리고 결과적인 쌍-말단이 서열화된다. 또한 Wei *et al.*, "next generation sequencing: towards personalised medicine. Michael Janitz Ed., 2008, Wiley section 13.2, Fig 13.1 보면 상기 서열분석은 부분적으로 더 긴 단편(대략 >1000bp)를 분석하기 위해 유용하다.

[0030] 타입 IIs 제한 엔도뉴클레아제(Type-IIs restriction endonuclease)는 제한효소 부위로부터 먼 인식 서열을 가지는 엔도뉴클레아제이다. 다시 말해, 타입 IIs 제한 엔도뉴클레아제들은 한 쪽에서 인식 서열의 바깥쪽을 절단한다. 상기의 예들은 NmeAIII(GCCGAG(21/19) 및 FokI, AlwI, MmeI이다. 양 쪽에서 인식 서열의 양쪽을 자르는 타입 IIs 효소들도 있다.

[0031] 정렬(Aligning 및 alignment): 여기에서 정의한 "정렬(aligning)" 및 "정렬(alignment)"이라는 용어는 동일한 또는 유사한 뉴클레오티드들의 짧은 또는 긴 서열들의 존재에 기반된 둘 또는 그 이상의 뉴클레오티드 서열의 비교를 의미한다. 뉴클레오티드 서열들의 정렬을 위한 다양한 방법들이 당해 기술 분야에서 잘 알려져 있다.

[0032] 여기에 사용된, 풀링(Pooling)은 풀(pools) 내에 있는 다수의 샘플들(또는 인공 염색체들(chromosomes) 또는 클론들(clones) 또는 유전체들의 부분집합들 또는 복사할 수 있는 복잡도 감소된 유전체들)의 조합과 관련 있다. 상기 풀링은 하나의 샘플(예를 들어, 각각 10 개의 샘플들을 포함하고 있는, 10 개 풀 내에 있는 100 개 샘플들) 내에 있는 다수의 개인 샘플들의 단순 조합일 수 있지만, 또한 더 정교한 풀링 전략들이 사용될 수 있다. 상기 풀에서 상기 샘플들의 분배는 바람직하게 각 샘플들이 적어도 둘 또는 그 이상의 풀에 존재하는 것이다. 바람직하게, 상기 풀은 각 풀당 10개부터 10000개까지의 샘플들, 바람직하게 100개부터 1000개까지, 더 바람직하게 250개부터 750개까지의 샘플들을 포함한다. 각 풀당 샘플들의 수는 광범위하게 변화할 수 있고, 그리고 상기 변화는, 예를 들어, 유전체의 크기 또는 조사 하에 샘플들의 수와 관련된다. 전형적으로, 풀 또는 서브-풀(sub-pool)의 최대 크기는 예를 들어 식별자들의 세트에 의해, 풀에 있는 샘플들을 고유하게 확인하는 능력에 의해 통제된다. 상기 풀은 당해 기술 분야에서 잘 알려진 풀링 전략들에 기반되어 발생된다. 본 발명의 기술 분야에서 숙련된 사람은 유전체 크기, 샘플의 수 등과 같은 요인들에 기반된 최상의 풀링 전략을 선택할

수 있다. 결과적인 풀링 전략은 환경에 의존할 수 있고, 그리고 상기의 예들은 플레이트 풀링(plate pooling), 3D-풀링(3D-pooling), 6D-풀링(6D-pooling) 또는 복잡한 풀링과 같은 N-차원 풀링(N-dimensional pooling)이다. 많은 수의 풀의 조작을 가능하게 하기 위하여, 상기 풀은, 그들 차례에서, 슈퍼-풀(super-pools)(예를 들어 슈퍼-풀은 샘플들의 풀의 풀이다)에 결합될 수 있고, 또는 슈퍼-풀로 나눠질 수 있다. 풀링 전략들 및 이들 데콘볼루션(deconvolution)(예를 들어 하나 또는 그 이상의 풀 또는 서브풀에 있는 샘플의 알려진 관련된 지표(예를 들어 라벨 또는 식별자)의 존재의 검출에 의해 라이브러리(library)에 있는 개인 샘플의 올바른 확인)은 예를 들어 US 6975943 또는 Klein et al. in Genome Research, (2000), 10, 798-807에서 설명된다. 상기 풀링 전략은 바람직하게 상기 라이브러리에 있는 모든 샘플이 풀의 고유의 결합이 모든 샘플을 위해 만들어지는 그런 풀에 분배되는 것이다. 상기의 결과는 (서브)풀의 특정 결합이 고유하게 샘플을 확인하는 것이다.

[0033] 클러스터링(Clustering): 여기에 정의된 "클러스터링(clustering)"이라는 용어는 짧은 또는 긴 동일한 또는 비슷한 뉴클레오티드들의 서열의 존재에 기반된 둘 또는 그 이상의 뉴클레오티드 서열의 비교 및 짧은 (또는 긴) 동일한 또는 비슷한 뉴클레오티드들의 서열의 존재에 기반된 특정 최소 수준의 서열 상동성(homology)을 가지는 서열들을 서로 그룹화하는 것으로 의미된다.

[0034] 본 발명의 상세한 설명(Detailed description of the invention)

[0035] 첫 번째 측면에 있어서, 본 발명은 하기 단계를 포함하는, 하나 또는 그 이상의 다수의 샘플들에 있는 하나 또는 그 이상의 다형성(polymorphisms)의 동시 발견, 검출 및 유전자형 판별을 위한 방법과 관계가 있다:

[0036] (a) 하나 또는 그 이상의 또는 다수의 샘플들로부터 DNA를 제공하는 단계;

[0037] (b) 제한효소단편들(restriction fragments)를 생산하기 위한 적어도 하나의 제한 엔도뉴클레아제들(resrestrc endonucleases)를 가지고 DNA를 절단함으로써 샘플 DNA의 복잡도를 감소시키는 단계;

[0038] (c) 태그된(tagged) 제한효소단편들을 생산하기 위해 식별자 태그(indentifier tag)를 갖는 샘플의 상기 제한효소단편들을 제공하는 단계;

[0039] (d) 상기 태그된 제한효소단편들의 적어도 일부분을 쌍-말단 서열분석(paired-end sequencing)하는 단계; 및

[0040] (e) 상기 샘플들 간 다형성을 확인하는 단계.

[0041] 상기 복잡도 감소(compexity reduction)는 오직 하나 또는 그 이상의 제한효소들을 가지고 샘플로부터 DNA의 절단에 기반된 것일 수 있다. 특정 실시예에서, 둘 또는 그 이상의 제한효소들이 사용될 수 있다. 제한효소단편들에, 어탭더들(adapters)이 라이게이션될 수 있다. 상기 어탭터들은 제한효소단편들의 한 쪽 말단 또는 양쪽 말단들에 라이게이션될 수 있고 그리고 상기 어탭터들은 동일하거나 또는 다를 수 있다. 상기 제한효소단편들이 둘 또는 그 이상의 다른 제한효소들을 가지고 DNA를 절단함에 의해 획득될 때, 다른 어탭터들이 사용될 수 있다. 복잡도 감소는 추가적으로 제한효소단편들을 증폭함에 의해, 예를 들어 상기 어탭터들 또는 상기 어탭터의 일부분으로 향하게 되는 프라이머들(primers)을 사용함으로써 달성될 수 있다. 증폭에 사용되는 상기 프라이머들은 추가로 제한효소들의 인식 서열의 나머지들에 상보적인 부분들을 포함할 수 있다. 특정 실시예에서, 단편들의 복사할 수 있는 부분집합을 제공하기 위해 프라이머들의 적어도 하나의 3' 말단에 1 내지 10 개의 무작위하게 선택적인 뉴클레오티드들(nucleotides)이 추가된다는 점에서AFLP®(EP534858)와 같은 확립된 기술들이 사용될 수 있다. 다른 복잡도 감소 기술들은 또한 이들 기술들이 복사할 수 있는(reproducible) 한 가능하다. 복사할 수 있는(reproducible)은 이 측면에서 동일한 샘플이 두 번 복잡도 감소를 당할 때, 동일한 부분집합이 얻어지고 그리고 두 샘플 간 충분히 동일한 부분집합이 획득되는 것을 의미한다.

[0042] 태그된 제한효소단편들을 생산하기 위한 상기 식별자 태그(identifier tag)는 수많은 방법들로 제공될 수 있다. 상기 식별자 태그는 하기에 의해 제공될 수 있다:

[0043] - 태그된 어탭터 라이게이션된(adaptor-ligated) 제한효소단편들을 생산하기 위한 제한효소단편들에 태그된 어탭터들을 라이게이션하는 것;

[0044] 또는

[0045] 태그된 어탭터 라이게이션된 제한효소단편들을 생산하기 위한 상기 어탭터의 적어도 일부분에 상보적인 적어도 하나의 태그된 프라이머를 가지고 상기 어탭터 라이게이션된 제한효소단편을 증폭하는 것.

[0046] 상기 어탭터는 오직 식별자 태그로 구성될 수 있고 또는 상기 어탭터는 예를 들어 상기 태그된 제한효소단편들(일부분)의 선택을 가능하게 하기 위한, 예를 들어 상기 샘플의 복잡도를 감소시키기 위한, 예를 들어 어레이(array) 상에서, 추가 기능들을 포함할 수 있다.

[0047] 상기 식별자 태그는 태그가 유래된 샘플에 제한효소단편을 연결하는 샘플당 고유의 태그가 제공되는 한, 또한 별개의 단계, 어탭터 라이게이션, 증폭 또는 복잡도 감소 전 또는 후에 추가될 수 있다.

[0048] 상기 서열분석 단계는 바람직하게 메이트-페어(mate-pair) 서열분석을 포함한, 고효율 서열분석을 사용하여, 쌍-말단 서열분석을 사용하여 수행된다.

[0049] 본 발명의 바람직한 실시예에서, 제한효소단편들의 서열의 일부분들이 결정된다. 바람직하게, 상기 제한효소단편들의 양 말단의 서열이 결정되고 그리고 바람직하게, 동시에, 예를 들어 동일한 서열분석 작업에서 결정된다. 상기 서열들의 결정을 가능하게 하는 절차들(protocols)은 전형적으로 여기에 정의된 메이트-페어 서열분석을 포함한, 쌍-말단 서열분석으로서 (GA*II* 및 Roche 플랫폼들을 위해 나타내어 진다.

[0050] 전형적으로 메이트-페어 서열분석을 포함한, 쌍-말단 서열분석을 이용하여, 제한효소단편의 두 말단의 서열 정보가 획득된다. '다이태그(ditag)'라고 불리는 것을 이끌어, 제한효소단편의 양 말단들(식별자를 포함한, 첫 번째 리드(read) 및 두 번째 리드)으로부터의 서열 정보가 결합될 수 있다. 상기 다이태그는 바람직하게 식별자 태그를 사용하는 샘플들에 연결될 수 있는 첫 번째 및 두 번째 리드의 결합된 정보를 포함한다. 상기 식별자 태그는 바람직하게 첫 번째 리드와 연관되어(또는 포함되어) 있다. 다이태그의 생성은 인 실리코(*in silico*)에 의해 될 수 있다. 바람직한 실시예에서, 리드들 중 하나, 바람직하게 두 번째 리드는, 다이태그의 생성 이전에 역 상보화된다. 역 상보화된(Reverse complemented)은 이 측면에서 리드의 서열이 반대로(예를 들어 N1N2N3N4N5N6이 N6N5N4N3N2N1이 된다) 되는 것이다. 따라서, 더 구체적으로 상기 다이태그는:

[0051] ID-read1-read2(역 상보화된): IDIDIDIDM1M2M3M4M5M6N6N5N4N3N2N1

[0052] 또한 본 개념의 설명을 위한 도면 1을 참조한다. 한 부분은 반복적인 서열로부터 획득될 수 있지만, 상기 다이태그의 다른 부분은 유전체 서열의 또다른 부분으로부터 획득될 수 있고, 따라서 두 부분들의 고유의 결합을 만들기 위한 기회가 증가한다. 이것은 다른 방법으로 불가능한 서열들 사이의 다형성의 확인을 가능하게 한다. 현재 기술은 150 개의 뉴클레오티드들이 300개의 정보를 제공하는 뉴클레오티드들을 이끄는, 상기 단편의 양쪽으로부터 획득되는 것을 가능하게 한다. 이것은 철저하게 샘플당 고유의 결합된 단편들의 수를 증가시키고 그리고 이 때문에 확인되는 다형성의 수를 증가시킨다. 동일한 기술적 개념은 메이트-페어 서열분석을 포함한, 쌍 말단을 가능하게 하는 다른 서열분석 플랫폼들에서 수행될 수 있다.

[0053] 고효율 서열분석은 바람직하게 합성에 의한 서열분석, 전형적으로 차세대 염기서열분석(Next Generation Sequencing)으로서 나타내어 지는 Illumina (GA*II*, Hiseq, MiSeq) 또는 Roche GS FLX에 의해 제공되는 플랫폼들과 같은 (고체 캐리어 상에서) 파이로서열분석(pyrosequencing)에 기반된다. 또한 차세대 염기서열분석으로서 나타내어지는 기술들이 사용될 수 있다. 상기 예들은 라이게이션에 의한 서열분석, 혼성화 서열분석(hybridisation sequencing), 나노세공 서열분석(nanopore sequencing)(Oxford 나노세공 기술들 또는 NABsys (US20100096268, US 20100078325, US20090099786))에 기반되고 또는 Pacific Biosciences 및 Ion torrent (*Nature* 475, Pages: 348-352)에 의해 제공된다.

[0054] 서열 정보를 획득함으로써, 서열들은 식별자 태그에 근거하여 샘플당 배치된다. 상기 서열들을 클러스터링(또는 정렬(aligning))함에 의해, 다형성은 서열들 간 확인될 수 있고 그리고 이 때문에 샘플들 간 다형성이 확인될 수 있다. 이는 SNPs의 확인, SNPs의 검출 및 동시에 다수의 샘플들의 유전자형들의 결정을 이끈다. 클러스터링 또는 정렬(alignmnet)은 당해 기술 분야에서 고식적 기술들을 사용하여 수행될 수 있다.

[0055] 비교 목적들을 위한 서열들의 정렬 방법들은 당해 기술 분야에서 잘 알려져 있다. 다양한 프로그램들 및 정렬 알고리즘들(algorithms)은 하기에서 설명되어 있으며: Smith and Waterman (1981) Adv. Appl. Math. 2:482; Needleman and Wunsch (1970) J. MoI. Biol. 48:443; Pearson and Lipman (1988) Proc. Natl. Acad. Sci. USA 85:2444; Higgins and. Sharp (1988) Gene 73:237-244; Higgins and Sharp (1989) CABIOS 5:151-153; Corpet et al. (1988) Nucl. Acids Res. 16:10881-90; Huang et al. (1992) Computer Appl. in the Biosci. 8:155-65; and Pearson et al. (1994) Meth. MoI. Biol. 24:307-31, 이는 여기에서 참조에 의해 포함되어 있다. Altschul et al. (1994) Nature Genet. 6:119-29(여기에서 참조에 의해 포함된)는 서열 정렬 방법들 및 상동성 계산들의 구체적인 고려 사항을 제시한다.

[0056] 서열 분석 프로그램들 blastp, blastn, blastx, tblastn 및 tblastx와 함께 연결하는데 사용하기 위해, NCBI Basic Local Alignment Search Tool(BLAST) (Altschul et al., 1990 J Mol Biol. 5;215(3):403-10)은 National Center for Biological Information(NCBI, Bethesda, Md.) 및 인터넷을 포함한 다양한 자료들로부터 이용 가능하다. <http://www.ncbi.nlm.nih.gov/BLAST/>에 접속될 수 있다. 상기 프로그램을 사용하여 서열 유사성(identity)을 결정하는 방법의 설명은 <http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html>에서 이용 가능하다.

[0057] 전형적으로, 정렬은 어탭터들/프라이머 및/또는 식별자들을 위해 다듬어진 서열 데이터들 상에서, 예를 들어 오직 핵산 샘플로부터 유래한 단편들로부터의 서열 데이터만 사용하여 수행된다. 전형적으로, 획득된 상기 서열 데이터는 (예를 들어 샘플로부터의) 단편의- 유래를 확인하기 위해 사용되고, 상기 어탭터 및/또는 식별자로부터 획득된 상기 서열들은 상기 데이터로부터 제거되고 그리고 정렬은 이 다듬어진 세트 상에서 수행된다.

[0058] 본 발명의 방법의 실시예에 있어서, 샘플들의 유전자 DNA(genomic DNA)는 두 개의 제한효소들, *Eco*RI 및 *Mse*I에 의해 절단된고, 어탭터들은 상기 단편들에 라이게이션된다. AFLP 복잡도 감소는 (유전체의 복잡도에 의존하여) 적용될 수 있다. 결국, 상기 결과로 얻어진 단편들은 GA*II* 서열분석에 적절하도록 만들어지고 쌍-말단 방식(fashion)(각 방향당 76 개 뉴클레오티드들)에서 서열분석된다. 태그 정의 및 유전자형 분류(calling)을 위한 생물정보학 접근법이 수행되고 결과로 얻어진 데이터의 분석은 샘플들 간 다형성의 확인을 이끌어 낸다. 이하, 본 발명을 실시예에 의해 상세히 설명한다.

[0059] 본 기술은 여러 점으로 추가된 유용성을 가진다:

[0060] 제한효소단편의 고효율 서열분석에 기반된 생리학적 지도를 만드는데 사용되는 동일한 제한효소를 사용함으로써, 서열화된 태그들 및 결과로 얻어진 유전자형들은 쉽게 생리학적 지도와 연결될 수 있다.

[0061] 쌍-말단 서열분석(예를 들어 각 단편의 상기 *Eco*RI 및 *Mse*I 말단을 말하는, 제한효소단편들의 양 말단을 서열분석)을 적용함으로써, *Eco*RI 및 *Mse*I 태그들의 유일한 고유의 결합을 정렬함에 뒤이어, 복제된 지역에서 SNP 분류 및 유전자형 판별이 최대화될 수 있다.

[0062] AFLP를 통해 견고한 복잡도 감소를 적용하는 것은 다수의 샘플들을 풀링하는 것을 가능하게 한다. 그러므로, 특정 실시예에서, 복잡도가 감소된 샘플들은 서열분석 이전에 풀 내에서 모여진다.

[0063] 본 기술은 바람직하게 총 유전자 DNA를 필요로 한다.

[0064] 본 출원을 통해, 본 발명이 제시한 더 많은 선행기술을 설명하기 위해 다양한 참조들이 삽입구(parentheses)에 인용되었다. 본 명세서에 인용된 모든 특허 및 문헌 참조는 이로써 전부 참조에 의해 포함된다.

[0065] 상기 설명 및 도면들은 본 발명의 몇몇 실시예를 설명하기 위해 포함되지만, 이에 보호 범위에 한정되지 않는다. 본 공개로, 더 많은 실시예들이 당해 기술 분야의 숙련된 사람 및 선행 기술들 및 본 특허 공개와 확실한 결합인 숙련된 사람들에게 명확해질 수 있다. 본 발명의 내용이 하기 실시예에 한정되는 것은 아니다.

[0066] **실시예(Examples)**

[0067] 본 프로젝트(project)의 목표는 무작위 서열 기반 유전자형 판별(random sequence based genotyping, rSBG)의 상황(context)에서 쌍-말단 서열(paired-end sequence) 데이터의 분석을 위한 전략을 만들어 내는 것이었다. 쌍-말단(다이태그들(ditags)) vs. Arbidopsis의 단일-말단(single-end) 전략을 사용하여 데이터를 분석하는 것의 수행이 평가되고 그리고 비교되었다. 상기 목적을 위하여, 참고 서열들이 Illumina GAII NGS 플랫폼(platform)으로부터의 상기 서열 데이터들을 가지고 데노보(de novo) 조립 전략을 사용하여 만들어 졌다. 그 다음, 상기 Illumina 리드들(reads)은 상기 참조 서열들에 지도화(mapped)된다. 상기 맴핑(mapping) 결과들은 다음 SNPs의 존재를 위해 탐색된다.

[0068] Arabidopsis 데이터세트(dataset)의 유전자 물질(gnentic material)은 두 부모, 두 F1 개인들 및 역교배(back cross, BC) 개체로부터의 28 자손으로 구성되었다.

[0069] 상기 쌍-말단 리드들은 단일 "리드(read)"에 결합되는, 다이태그들로 불리는, 구조들을 만들기 위해 사용되었다. 상기 다이태그의 길이는 상기 쌍에서 각 리드의 길이의 합이다. 추가적으로, 다이태그를 만들기

전 read2 리드들이 참조(유전체(genome)) 서열에서 다이태그의 맵핑이 가능하게 하기 위해 역-상보화(reverse-complemented)되었다. 따라서, 상기 다이태그들의 최종 구조는 하기였다:ID tag-read1-read2(역-상보적). 상기 다이태그들은 품질 관리(quality control) 단계들의 어느 하나가 수행되기 전에 만들어졌고, 그리고 상기 품질 관리 절차들은 쌍-말단 서열 데이터로부터의 각 리드 파일의 필터링(filtering)에 사용되는 동일한 기준들(criteria)를 가지고 상기 다이태그들을 걸러내기 위해 적용되었다.

[0070] 상기 ID tag는 쌍-말단 서열 데이터로부터의 read1 및 read2 서열들 모두에 존재하였다.

[0071] EcoRI/MseI 라이브러리들은 상기 Arabidopsis 샘플들의 각각을 위해 발생되고 그리고 상기 Illumina GAII를 사용하여 서열분석된다. 품질 관리 접근법들은 다이태그들 및 상기 쌍-말단 서열 데이터로부터 획득된 상기 read1 및 read2 파일들을 위해 수행된다.

[0072] 상기 Arabidopsis 서열 데이터에 적용된 품질 관리 필터링을 위한 요약 통계는 [표 1]에 나타내었다.

표 1

[0073] Arabidopsis에 있는, Illumina GAII 서열 데이터를 위한 기술적인 통계

	다이태그(Ditags)	Read1	Read2
리드들(reads)의 초기 수	19,622,319	19,622,319	19,622,319
ID tags 없는 리드들	594,273	594,273	594,273
*Eco*RI 제한효소 패턴 없는 리드	3,136,557	3,136,557	n.a.
*Mse*I 제한효소 패턴[§] 없는 리드들	1,495,914	n.a.	4,390,806
동종 중합체(homopolymer) 서열들을 포함한 리드들	18,298	39,849	17,766
엽록체/미토콘드리아 데이터베이스에 상당한 일치를 가지는 리드들	3,438,320	3,704,597	3,399,068
미결정의 뉴클레오티드들(nucleotides)을 포함하는 리드들	25,632	32,177	23,621
저품질의 리드들	32,452	54,647	138,345
필터링 후 리드들의 최종 수	10,880,838	12,060,184	11,058,405
QC를 통과한 ID tags를 가진 리드들의 %	55.5	61.5	56.4

[0074] [§]. 다이태그들 품질 관리에 있어서, *Mse*I 패턴의 존재는 다이태그의 말단에서 평가되었다.

[0075] 상기 서열분석 레인(lane)에서 생산되는 리드들의 총 수는 19,622,319였다. 총 상기 리드들의 97%가 초기에 ID tag를 가졌고, 이는 샘플들과도 일치될 수 없는 리드들 때문에 상기 서열들의 유일하게 작은 퍼센트가 제거되었음을 나타낸다. 모든 필터링 기준의 적용 후, 데이터세트에 남아있는 리드들의 수는 10.9M(다이태그들)부터 12.1M(read1)까지의 범위였다. 상기 데이터세트로부터 리드들의 제거의 주된 이유들은 기대되는 제한효소 모티프(motif)(EcoRI 또는 MseI)의 부재, 및 엽록체/미토콘드리아 데이터베이스에서 상당한 일치를 가지는 리드들 때문이다.

[0076] 다이태그 vs. Arabidopsis에 있는 단일-말단의 비교는 다이태그 또는 단일-말단으로서 데이터를 분석하는 것의 수행을 평가하기 위하여, 고유의 리드들을 이용한 CAP3 조립들(Huang *et al.* Genome Res. 1999 Sep;9(9):868-77)을 이용하여 만들어졌다. 상기 단일-말단 분석은 별도로 쌍-말단 서열 데이터의 각 리드 파일을 위해 수행되었고, 그리고 최종 결과들은 각 리드 파일과 함께 분석하여 얻어진 수들을 추가함으로써 결정되었다. 본 평가의 요약 결과들은 [표 3]에 나타내었다.

표 3

[0077] Arabidopsis 서열 데이터세트(CAP3 및 고유의 리드들을 가지고 수행된 조립들)에서, 조립 전략의 비교(다이태그 vs. 단일-말단) 요약 결과들

	다이태그들	단일-말단 read1	단일-말단 read2	결합된 단일-말단
리드들의 수	689,512	597,878	784,782	1,382,660
콘틱들(contigs)의 수	31,891	38,236	43,448	81,684
SNPs를 가지는 콘틱들	4,371	1,956	1,774	3,730
SNPs 수[§]	8,760	3,338	2,838	6,176
콘틱당 SNPs 수	2.0	1.7	1.6	1.7
SNPs 수[¶](90% 유전자형 판별 비율)	2,634	1,385	997	2,382
유전자형들 수(90% 유전자형 판별 비율)	78,174	41,063	29,533	70,596
SNPs 수[†](80% 유전자형 판별 비율)	3,346	1,791	1,352	3,143
유전자형들의 수(80% 유전자형 판별 비율)	96,779	51,645	38,808	90,453

[0078] [§]엄격한 세팅(settings)이 없는 상태에서 확인된 SNPs 수 [¶]적어도 28 개인들에서 유전자형이 판별되는 90% 유전자형 판별 비율 [†]적어도 25 개인들에서 유전자형이 판별되는 80% 유전자형 판별 비율

[0079]

[0080] 두 유전자형 판별 근처에서, SNPs 및 유전자형들의 수는, 상기 데이터가 다이태그들로서 분석될 때 더 높았다. 상기 증가된 수행으로 각각, 추가적인 11% 및 7% SNPs 및 90% 및 80% 유전자형 판별 비율에서의 유전자형들의 확인 결과를 얻었다. 그 다음, 상기 Arabidopsis 데이터에 이용할 수 있는 역교배 개체 구조를 이용하여, 각 SNP 데이터세트를 위한 A, B 및 H 유전자형의 수가 결정되었다. 상기 개체의 역교배 본성 때문에, 오직 한 동형 접합체(homozygote) 유전자형이 관찰되었고, 따라서 B 유전자형의 수는 유전자형 분류(calling)에서 전체 오류의 좋은 지표이다. 추가적으로, A 및 H 유전자형들의 빈도는 대략 50% 정도여야 하고, 그리고 상기 빈도로부터 큰 편차들 또한 유전자형 분류가 가지는 문제들의 신호이다. Arabidopsis 데이터에서 수행된 유전자형 검사의 결과들은 [표 4]에 포함되어 있다.

[0081]

표 4

[0082] Arabidopsis 데이터에서 수행된 유전자형 검사 결과

	다이태그들(ditags)	Read1	Read2
SNPs 수	2,334	1,101	1,024
유전자형들 수	58,452	27,647	25,078
A 유전자형	25,108	11,911	10,841
% A	43.0	43.1	43.2
B 유전자형	447	252	156
% B	0.8	0.9	0.6
H 유전자형	32,897	15,484	14,081
% H	56.3	56.0	56.1

[0083] 상기 유전자형 분류의 정확도는 오직 부모들이 대체 대립유전자(alternate alleles)에서 동형 접합체인 SNPs를 위해 수행된다. 상기 결과들은 B 유전자형의 빈도가 모든 검사된 전략들에서 1% 이하였기 때문에, 유전자형 판별 정확도가 높음을 확실하게 한다. 뿐만 아니라, 본 발명은 각 유전자형 부류에 대한 빈도들이 검사된 모든 전략들과 현저히 유사하기 때문에, 검사된 상기 세 가지 분석 전략들 간 유전자형 판별 정확도에서 상당한 차이가 없음을 확인하였다.

[0084] 상기 결과들은 SNPs 분류 및 유전자형 판별의 정확도를 포함하는 것 없이, 상기 다이태그 분석이 수많은 SNPs

및 유전자형들을 발생시킨다는 것을 확인하였다.

도면

도면1