

명세서

발명의 명칭: 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법 및 예측 시스템

기술분야

- [1] 이하 설명하는 기술은 유전자 발현 데이터를 이용하여 유방암 예후를 예측하는 기법에 관한 것이다.

배경기술

- [2] 종양의 예후를 예측하기 위한 다양한 연구가 진행되고 있다.
- [3] 예컨대, 유방암(breast cancer) 분야는 유전자 분석을 통해 유방암 예후를 예측하는 기술이 개발되고 있다. 대표적으로 유방암 예후예측 도구의 표준으로 사용되고 있는 Oncotype DX®, Mammaprint® 등은 모두 RT-PCR(Real-time PCR) 기반의 검사이다.

발명의 상세한 설명

기술적 과제

- [4] RT-PCR 기반의 검사는 비용과 효율성의 측면에서 많은 유전자를 동시에 분석하기 어려운 한계가 있다. 이하 설명하는 기술은 NGS(next-generation sequencing) 기법으로 유전자 발현량을 분석하여 유방암 예후를 추정할 수 있는 기법을 제공하고자 한다.

과제 해결 수단

- [5] 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법은 컴퓨터 장치가 피험자 조직(tissue)의 RNA 시퀀싱 데이터를 이용하여 타겟 유전자의 발현량을 측정하는 단계, 상기 컴퓨터 장치가 상기 타겟 유전자의 발현량을 사전에 마련한 인공신경망(Artificial Neural Network)에 입력하는 단계 및 상기 컴퓨터 장치가 상기 인공신경망의 출력값을 기준으로 상기 피험자에 대한 유방암 예후를 추정하는 단계를 포함한다.
- [6] 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 시스템은 피험자 조직(tissue)의 RNA 유전자 데이터를 저장하는 클라이언트 장치 및 상기 유전자 데이터에 대한 NGS(next-generation sequencing) 기반 시퀀싱을 수행한 RNA 시퀀싱 데이터를 이용하여 타겟 유전자의 발현량을 측정하고, 사전에 마련한 인공신경망(neural network)에 상기 타겟 유전자의 발현량을 입력하여 출력되는 출력값을 기준으로 상기 피험자에 대한 유방암 예후를 추정하는 분석 서버를 포함한다.
- [7] 상기 인공신경망은 복수의 샘플의 타겟 유전자 발현량을 입력값으로 갖고, 상기 복수의 샘플에 대한 온코타입(Oncotype DX)의 재발 점수(recurrencescore)에 기반한 결과를 출력하도록 사전에 마련된다.

발명의 효과

- [8] 이하 설명하는 기술은 NGS 기법을 이용하여 RT-PCR 기반 기법에 비하여 타겟 유전자 서열을 보다 저렴한 비용으로 빠르게 분석할 수 있다. 이하 설명하는 기술은 타겟 유전자의 발현량과 Oncotype DX의 재발 점수(recurrence score)로 학습한 기계학습(machine learning) 모델을 이용하여 빠르면서도 정확한 유방암 예후 예측이 가능하다.

도면의 간단한 설명

- [9] 도 1은 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법에 대한 순서도의 예이다.
- [10] 도 2는 타겟 유전자에 대한 예이다.
- [11] 도 3은 NGS 기반한 타겟 유전자의 데이터 품질에 대한 예이다.
- [12] 도 4는 특정 타겟 엑손 영역의 유전자 발현 데이터를 이용하는 예이다.
- [13] 도 5는 특정 타겟 엑손 영역을 활용한 방식의 안정성을 검증한 예이다.
- [14] 도 6은 RNA 시퀀스 데이터를 정규화하는 과정에 대한 예이다.
- [15] 도 7은 기계학습의 인공신경망 구조에 대한 예이다.
- [16] 도 8은 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 시스템에 대한 예이다.

발명의 실시를 위한 형태

- [17] 이하 설명하는 기술은 유전자 발현 데이터를 이용하여 유방암 예후를 예측하는 기법에 관한 것이다. 유방암 예후를 예측하는 도구로 기계학습모델을 사용한다. 기계학습모델은 유전자 발현량과 Oncotype DX®(이하 Oncotype DX라 함)의 재발 점수(recurrence score, 이하 RS라 함)를 이용하여 학습한다. 먼저 Oncotype DX에 대하여 간략하게 설명한다.
- [18] Oncotype DX는 미국 Genomic Health 사에서 개발한 분석 툴이다. Oncotype DX는 유방암 조직에서 21개의 다른 유전자의 활성도를 측정하고 분석하여 유방암이 재발한 가능성 및 화학요법이 얼마나 효과가 있을지 알려주는 검사이다.
- [19] Oncotype DX는 16개의 유전자와 5개의 레퍼런스(reference) 유전자를 기준으로 RS를 계산한다. RS를 계산하는 수식은 다음과 같다. $RS = +0.47 \times \text{HER2 Group Score} - 0.34 \times \text{ER(Estrogen) Group Score} + 1.04 \times \text{Proliferation Group Score} + 0.10 \times \text{Invasion Group Score} + 0.05 \times \text{CD68} - 0.08 \times \text{GSTM1} - 0.07 \times \text{BAG1}$ 이다. 여기서 각 항목은 기능적 구분에 따른 유전자 그룹을 말한다. HER2 Group(2개 유전자), ER Group(4개 유전자), Proliferation Group(5개 유전자) 및 Invasion Group(2개의 유전자)은 복수의 유전자를 포함한다. CD68, GSTM1 및 BAG1는 각각 하나의 개별 유전자이다. RS 점수는 0 ~ 100점 범위의 값을 갖는다. 예컨대, RS 점수가 낮을 경우 재발율이 낮고, 화학치료의 효과도 낮을 가능성이 높다. 반대로 RS 점수가 높다면 재발율이 높고, 화학치료가 효과적일 가능성이 높다.

- [20] 전술한 바와 같이 Oncotype DX는 RT-PCR에 기반하여 유전자 활성도를 검출하지만, 이하 설명하는 기술은 NGS 기법을 사용한다. 이하 설명하는 기술은 NGS 기법에 기반하여 RNA를 시퀀싱하고 유전자 발현량을 결정한다. NGS 기법에 대해서도 간략하게 설명한다.
- [21] NGS 기술은 수십만개의 반응을 동시에 수행하는 멀티플렉싱(multiplexing) 능력이 있으며, 적은 양의 샘플로도 시퀀싱이 가능하다. NGS는 상용화된 기술에 따라 구체적인 적용 기법이 다소 다르지만, 일반적으로 클론증폭(clonal amplification), 대량병렬 시퀀싱 및 Sanger 방법과 작용기전이 다른 새로운 염기서열결정법을 사용한다. 상용화 기술을 간략하게 소개한다. Roche사는 2007년 454 Cooperation사를 454 GS 개량형 FLX model sequencer를 출시하였다. Illumina사는 2006년 Genome Analyzer HiSeq를 출시하였고, Applied Biosystems사는 2007년 SOLiD를 차례로 출시하였다. 세 가지의 플랫폼은 공통적으로 복잡한 라이브러리 구축과 클로닝과정을 버리고 클론증폭기술을 채택하였고, 한꺼번에 대량으로 처리할 수 있는 대량병렬방식(massively parallel sequencing) 기술을 택하였으며, 순환 시퀀싱(cyclic sequencing)을 통한 합성신호읽기(sequencing by synthesis)로 염기서열을 결정하여 번잡한 전기영동과정을 배제하였다. 또한, shotgun 방식을 사용하여 읽혀진 짧은 리드(read)를 컴퓨터로 배열하여 중복된 부분을 찾아 전체를 완성하는 알고리즘을 사용한다.
- [22] 이하 설명하는 기술은 특정 조직(유방 조직)에서 추출한 샘플에 대한 유전자 발현량(RNA 발현량)을 기준으로 유방암 여부를 예측한다. 이하 특정 조직에서 샘플을 준비하고 RNA를 추출하는 과정에 대한 예를 먼저 설명한다.
- [23] **대상 유방암 환자 선정 및 검사 조직 준비**
- [24] 1) 호르몬 수용체 양성, 임파선 전이 음성인 1-2기 유방암의 수술 조직 중 대표 포르말린고정파라핀포매 (formalin-fixed paraffin-embedded, FFPE)블록을 선정한다.
- [25] - 병리전문의의 H&E 염색 슬라이드 확인 하 블록을 선정함. 이때 대상이 되는 종양이 확실히 존재하는 블록을 선택하여야 하며 가급적 한 단면 내 종양의 면적이 가장 넓은 것, 종양조직 내 괴사부가 적거나 존재하지 않는 것이 좋다.
- [26] 2) 10 μ m 두께의 비염색 슬라이드 10장을 준비한다.
- [27] **FFPE 조직으로부터 RNA 추출 프로토콜**
- [28] 1) RNA 추출 키트: 아래의 상용화된 두 키트 중 한 가지를 사용 할 수 있다.
- [29] ① Ambion RecoverAll™ Total Nucleic Acid Isolation Kit for FFPE
- [30] ② QIAGEN RNeasy FFPE Kit
- [31] 2) Wash 용액 준비
- [32] - Wash 1에 42mL의 100mL 에탄올을 섞는다 -> Wash 1
- [33] - Wash 2/3에 48mL의 100mL 에탄올을 섞는다 -> Wash 2/3
- [34] 3) 탈파라핀 (준비 : 조직, 100% 자일렌, 100% 에탄올, 가열 블록 50°C, 피펫,

볼텍스 믹서, 원심분리기)

- [35] ① 조직 준비 : 파라핀 블록에서 깎아낸 10um 두께의 파라핀 절편 4~8장, 총 40-80um을 준비한다. 이 때 절편 내 종양의 크기가 40mm보다 작다면 8장의 절편 모두를 사용한다. 가급적 절편 중 종양부만 들어가도록 유의하며 준비된 절편을 1.5mL tube에 담는다.
- [36] ② 1mL의 100% 자일렌을 조직에 넣고 볼텍스 믹서로 섞고 짧게 원심 분리한다. 이후 3분동안 50°C에 두어 파라핀이 녹도록 한다. (녹지 않으면 이 과정을 다시 반복한다)
- [37] ③ 2분동안 최대속도로 원심분리 하여 덩어리가 만들어지도록 한다. 단단히 뭉쳐지지 않으면 2분의 원심분리를 추가한다. 덩어리를 망가트리지 않고 자일렌을 버린다.
- [38] ④ 자일렌 씻어내기
- [39] a. 1mL의 100% 에탄올을 검체에 넣고 볼텍스 믹서로 섞는다. (뿌영게 된다.)
- [40] b. 1분간 상온에서 최대속도로 원심분리하여 덩어리가 만들어지도록 한다.
- [41] c. 덩어리가 망가지지 않도록 하면서 에탄올을 제거한다.
- [42] d. a-c과정을 한 번 반복한다.
- [43] e. 짧게 원심분리 후 가능한 한 덩어리를 건드리지 않으면서 남아있는 에탄올을 최대한 제거한다.
- [44] ⑤ 상온에서 15-45분간 건조시킨다.
- [45] 4) 단백질 분해 (준비 : 가열 블록 50°C & 80°C, Protease(단백분해효소)는 냉동실에서 꺼내서 상온에서 녹여둔다)
- [46] ① Digestion Buffer 200 μ L와 Protease 4 μ L를 각 검체에 넣는다. 이때 잘 섞이도록 부드럽게 흔들어준다.
- [47] ② 검체를 50°C(protease 활성화 온도) 가열 블록에 15분 이상, 완전히 투명해 질 때까지 둔다.
- [48] ③ 이후 80°C(protease 비활성화 온도) 가열 블록에 15분동안 둔다. 이때, 시간을 정확히 지킨다.
- [49] ④ 녹지 않으면 protease만 4 μ L 추가하여 위 과정(2 & 3)을 반복한다.
- [50] 5) 핵산 분리 (준비 : Isolation Additive/에탄올 혼합물, 기타 시약 모두)
- [51] ① Isolation Additive/에탄올 혼합물 제조
- [52] - Isolation Additive 240 μ L + 100% 에탄올 500 μ L = 총 790 μ L
- [53] - 제조 후 50mL 튜브에 보관한다.
- [54] (다수의 검체를 대상으로 할 때는 정량보다 5% 정도 많게 준비하도록 한다.)
- [55] ② 제조한 Isolation Additive/에탄올 혼합물을 790 μ L씩 각 검체가 담긴 튜브에 분주하면서 피펫을 이용해 섞는다.
- [56] ③ 혼합물 거르기
- [57] a. 필터 카트리지를 키트에서 제공된 튜브에 넣는다.
- [58] b. 2번 과정에서 만든 혼합물 700 μ L를 필터 위에 올리고 뚜껑을 닫는다.

- [59] c. 10,000rpm에서 30초간 원심분리한다.
- [60] d. 걸러져 나온 용액은 버리고 필터를 같은 튜브에 넣는다.
- [61] e. 필요한 경우(혼합물이 충분히 걸러지지 않은 경우) 원심분리를 한 번 더 실시해 혼합물이 필터에 걸러지도록 한다.
- [62] ④ Wash 1
- [63] a. 700 μ l의 Wash 1을 필터 카트리지에 더한다
- [64] b. 10,000rpm에서 30초간 원심분리한다.
- [65] c. 걸러져 나온 용액은 버리고 필터를 같은 튜브에 넣는다.
- [66] ⑤ Wash 2/3
- [67] a. 500 μ l의 Wash 1을 필터 카트리지에 더한다
- [68] b. 10,000rpm에서 30초간 원심분리한다.
- [69] c. 걸러져 나온 용액은 버리고 필터를 같은 튜브에 넣는다.
- [70] d. 10,000rpm으로 한번 더 원심분리하여 남은 용액을 제거한다.
- [71] 6) RNA 분리 및 정제 (준비 : DNase(DNA 분해효소)와 Nuclease(핵산분해효소)는 냉동실에서 꺼내서 녹여둔다)
- [72] ① RNA 분리
- [73] a. DNase 혼합물 제조: 10X DNase Buffer 6 μ l + DNase 4 μ l +Nuclease free water 50 μ l = 총 60 μ l
- [74] b. DNase 혼합물 60 μ l를 각각의 필터 카트리지 중앙에 더한다.
- [75] c. 뚜껑을 닫고 22-25℃의 상온에서 30분간 둔다.
- [76] ② Wash 1
- [77] a. Wash 1 700 μ l를 필터 카트리지에 넣고 30-60초간 상온에 둔다.
- [78] b. 10,000rpm에서 30초간 원심분리한다.
- [79] c. 걸러져 나온 용액은 버리고 필터를 같은 튜브에 넣는다.
- [80] ③ Wash 2/3
- [81] a. Wash 2/3 500 μ l를 필터 카트리지에 넣는다.
- [82] b. 10,000rpm에서 30초간 원심분리한다.
- [83] c. 걸러져 나온 용액은 버리고 필터를 같은 튜브에 넣는다.
- [84] d. a-c를 한번 더 반복한다.
- [85] e. 10,000rpm에서 1분간 원심분리한다.
- [86] ④ Elution solution 더하기 및 보관
- [87] a. 필터 카트리지를 새 튜브에 넣는다.
- [88] b. 60 μ l Elution Solution을 필터의 중앙에 더한다.
- [89] c. 뚜껑을 닫고 1분간 둔다.
- [90] d. 1분간 최대속도로 원심분리 한 후 필터는 버리고, 걸러져 나온 용액을 -20℃ 이하에서 보관한다.
- [91] 이하 샘플 조직에서 추출한 RNA를 이용하여 유방암 예후를 예측하는 과정에 대하여 설명한다. 도 1은 차세대 염기서열분석을 이용한 기계학습 기반 유방암

예후 예측 방법에 대한 순서도의 예이다. 먼저 RNA 샘플에 대한 시퀀싱을 수행한다(110). 다양한 방법으로 RNA 시퀀싱을 수행할 수 있다. 다양한 상용 키트(kit) 및 상용 솔루션을 사용하여 RNA를 시퀀싱할 수 있다. 하나의 예를 설명한다.

- [92] **Targeted RNA-sequencing**
- [93] 1) KAPA Stranded RNA-Seq kit with RiboErase (KK8483, KAPABIOSYSTEMS) 키트를 사용하여, 전체 RNA에서 ribosomal RNA를 제거한다.
- [94] 2) mRNA로부터 cDNA를 제작하고 추가적인 과정을 통해 cDNA NGS 라이브러리(Library)를 완성한다. cDNA Library 및 Hybridization solution, Target Capture Probe를 이용하여, 액상 혼성화 유전자 포획 기법 (Solution-based hybridization capture)을 진행한다.
- [95] 3) 획득된 산물을 일정 부분 증폭하여 라이브러리 증폭(Library amplification)을 진행한다.
- [96] 4) 최종 산물을 시퀀싱(Illumina사의 키트 사용)하여 생성된 타겟 영역의 시퀀싱 깊이(depth) 데이터를 기반으로 RNA의 발현량을 예측할 수 있다.
- [97] NGS 기반의 RNA 시퀀싱을 수행하면 상용 프로그램에서 RNA 시퀀스를 일정한 디지털 데이터로 생성한다. 또한 상용 프로그램은 시퀀싱 결과를 이용하여 각 RNA 유전자의 발현량을 산출할 수 있다. 따라서 RNA 샘플을 분석하는 과정 및 분석 결과는 컴퓨터 장치가 수행한다. 따라서 이하 컴퓨터 장치가 RNA 발현량을 이용한 유방암 예후 예측을 수행한다고 전제한다. 컴퓨터 장치는 샘플 RNA에 대한 발현량 데이터를 생성한다(120).
- [98] 한편 샘플 RNA 전체 유전자에 대하여 분석을 수행하지 않고, 유방암 예후와 연관성 있는 유전자(이하 타겟 유전자라함)를 선별하여 분석을 수행할 수 있다. 타겟 유전자를 결정하기 위하여 공개된 유전자 데이터를 활용하였다. 에스트로겐 수용체가 양성이고, 임파선에 전이가 되지 않은 피험자의 유전자 데이터를 활용하였다. 활용한 공개 데이터는 GSE2034, GSE2990, GSE3494, GSE4922, GSE6532, GSE7390 및 GSE12093이다.
- [99] 공개 데이터를 기반으로 각 유전자와 Oncotype DX RS(재발 점수)의 상관 관계(correlation)를 분석하였다. 대표적인 상관 관계 분석 기법인 피어슨(Pearson)와 스피어만(Spearman) 기법을 활용하였다. 각 공개 데이터에서 Oncotype DX RS(재발 점수)와의 상관 계수가 평균 0.5 이상인 유전자를 선별하였다. 아래 표와 같이 모두 135개의 유전자를 선별하였다.

[100] [표1]

	GSE20 34	GSE29 90	GSE34 94	GSE49 22	GSE65 32	GSE73 90	GSE120 93	합계(>0.5)
Oncotype DX RS	184	298	267	292	168	77	56	135

- [101] 또한 추가적으로 Oncotype DX RS 계산에 사용되는 16개의 유전자를 추가 선별하였다. 결국 타겟 유전자는 149개의 유전자를 활용하였다. 물론 실험적으로 타겟 유전자는 다른 유전자 조합을 선택할 수도 있을 것이다. 다만 타겟 유전자는 Oncotype DX RS와의 상관도가 큰 유전자로 결정한다.
- [102] 도 2는 타겟 유전자에 대한 예이다. 도 2는 모두 149개의 유전자를 표시한다. 도 2에서 음영으로 표시한 유전자는 세포 분열을 조절하는 기작(Cell Cycle)과 관련된 유전자이다. 도 2에서 실선 원으로 표시한 유전자는 p53 신호 경로에 관여하여 세포 분열을 조절하는 기작과 관련된 유전자이다. 도 2에서 점선 원으로 표시한 유전자는 DNA 복제 과정을 조절하는 유전자이다. 도 2에서 실선 사각형으로 표시한 유전자는 Cell cycle 및 p53 signaling pathway에 관여하는 유전자이다. 도 2에서 점선 사각형으로 표시한 유전자는 Cell cycle 및 DNA 복제에 관여하는 유전자이다.
- [103] 전술한 바와 같이 NGS 기반하여 컴퓨터 장치는 타겟 유전자의 발현량을 측정한다. 도 3은 NGS 기반한 타겟 유전자의 데이터 품질에 대한 예이다. 도 3은 타겟 RNA 시퀀싱한 데이터의 품질을 나타내는 그래프의 예이다. 데이터 품질 검증을 위하여 측정된 유전자 발현량 값과 전체-전사체 시퀀싱으로 측정된 값 사이의 편차를 확인한 것이다. 총 84개의 패널 유전자를 대상으로 10개의 RNA 샘플에 대해 표적 시퀀싱과 전체-전사체 시퀀싱으로 발현량을 각각 산출하고, 피어슨 상관 계수를 측정하였다. 결과적으로 0.85이상의 높은 상관 관계를 확인하였다. 따라서, 타겟 RNA 시퀀싱과 전체-전사체 RNA 시퀀싱 결과가 높은 상관 관계를 보이므로, 실험에서 사용한 타겟 RNA 시퀀싱의 결과는 전체-전사체 RNA 시퀀싱과 유사한 정도의 데이터 품질을 갖는다고 해석된다.
- [104] 샘플에 대한 전체 유전자 발현량을 산출하지 않고, 각 샘플에서 공통되게 발현하는 특정 엑손(exon) 부위를 사용할 수 있다. 이 경우 유전자 발현량 연산 과정이 보다 빠르게 수행될 수 있다. 즉 샘플에 대하여 공통되게 보존된 엑손(conserved exon) 패널을 이용하는 것이다(이하 CE 방식이라고 함).
- [105] 도 4는 특정 타겟 엑손 영역의 유전자 발현 데이터를 이용하는 예이다. 도 4는 복수의 샘플(A 내지 D)에 대한 유전자 엑손의 발현 상태를 나타낸다.
- [106] 복수의 샘플에 대하여 공통되게 발현하는 영역(타겟 영역)을 결정하여 해당 영역을 기준으로 유전자 발현량을 결정할 수 있다. 유전자 전체 영역을 이용하는 방식(이하 WG 방식이라고 함)은 개개인의 유전자 아형 전사체 발현량 차이가

고려되지 않기 때문에 환자별 아형 전사체 발현 비율 차이로 인해 유전자 발현량 측정에 편차가 발생할 수 있다. CE 방식은 아형 전사체들이 공유하는 영역만을 사용하기에 WG 방식에서 발생하는 측정편차가 발생하지 않아 보다 안정적으로 발현량 측정이 가능하다. 안정성은 반복 측정시에 측정값이 변하는 정도를 의미한다.

- [107] 나아가 샘플 데이터를 활용하여 CE 방식의 안정성을 검증하였다. 도 5는 특정 타겟 엑손 영역을 활용한 방식의 안정성을 검증한 예이다. 도 5(A)는 CE 방식의 안정성 검증을 위한 과정을 설명하는 예이다. Replicate 데이터는 단일 샘플에 대하여 반복측정한 값이다. 따라서 반복 측정한 데이터 사이의 오차는 측정 오차에 해당한다. 중앙 셀 라인에서 추출한 9개의 RNA 샘플에 대해 각 2회씩의 반복 실험을 수행하여 총 18개의 표적 RNA 샘플 시퀀싱 데이터를 생성하였다(타겟 RNA 시퀀싱 I 및 타겟 RNA 시퀀싱 II). 도 5(B)는 WG 방식의 측정값과 CE 방식의 측정값 사이의 안정성을 나타내는 실험 결과이다. 총 9개의 샘플 중 8개 샘플에서 WG 방식 보다 CE 방식의 Replicate 데이터간 비교 시에 더 안정적이라고 나타난다(피어슨 계수가 더 높다). 9개의 샘플에 대한 평균 피어슨 계수도 CE 방식이 WG 방식보다 높고, 샘플간 표준편차도 적다. 따라서 CE 방식이 WG 방식보다 더 안정적이라고 해석된다.
- [108] 도 1에 대한 설명으로 돌아가면, 컴퓨터 장치는 생성한 타겟 유전자의 발현량 데이터를 일정하게 후처리하고 정규화할 수 있다(130). 먼저 적용 가능한 후처리 과정의 예를 설명한다. 후처리 및 정규화는 디지털 데이터(파일)을 일정하게 처리하는 과정에 해당한다.
- [109] **Targeted RNA-sequencing 결과 후처리**
- [110] 1) 리드(read) 제거
- [111] 리드 품질 기준을 만족하지 않는 리드들을 제거한다(예컨대, 평균 품질이 20 이상이고, 평균 품질이 2 이하 염기(base)가 5% 미만되도록 처리함)
- [112] Trimmomatic (0.33) 프로그램을 사용하여 시퀀싱 과정에서 삽입된 인덱스 시퀀스를 제거한다.
- [113] 2) 시퀀싱된 리드를 레퍼런스 게놈에 정렬(Align sequenced reads to the reference genome)
- [114] STAR aligner 프로그램을 사용하여 시퀀싱된 리드들의 레퍼런스 게놈(hg19)을 기준으로 한 위치를 찾고 SortedByCoordinate 옵션을 주어 동시에 정렬한다.
- [115] 3) 유전자 발현량 연산
- [116] cufflinks 프로그램을 사용하여 정렬된 리드 정보들로부터 유전자별 발현량과 전사물별 발현량을 계산한다. 발현량은 FPKM(Fragments Per Kilobase of exon per Million fragments mapped)값으로 계산할 수 있다. 유전자별로 계산된 발현량은 genes.fpkm_tracking 파일로 생성되고, 전사물별로 계산된 발현량은 isoforms.fpkm_tracking 파일로 생성될 수 있다.
- [117] 한편 정확한 시퀀스 정렬 및 발현량 측정을 위해 분석 전 처리 과정으로 시퀀싱

결과로 나온 리드 들 중 염기 품질이 좋지 않은 리드들을 제거하고 각 리드들의 말단부에 남아있을 수 있는 (시퀀싱 과정에서 삽입된) 인덱스 시퀀스를 제거할 수 있다. 전 처리 과정을 거친 리드들에 대해서 STAR 프로그램을 사용하여 각각의 리드들의 레퍼런스 게놈 상의 위치를 확인. 확인된 위치정보는 BAM 파일 포맷으로 생성되고 이 BAM 파일을 Cufflinks 프로그램을 이용하여 유전자 및 전사물 별 발현량을 계산할 수 있다.

- [118] 컴퓨터 장치는 생성한 데이터를 정규화할 수 있다(130).
- [119] **Targeted RNA-sequencing 발현정보 정규화 (normalization)**
- [120] 종래 정규화 기법들 중 R package edgeR (Robinson et al. Bioinformatics 2010)에서 사용되는 "Trimmed Mean of M-value (TMM)" 기법이 가장 안정성이 높은 것으로 알려져 있다. 컴퓨터 장치는 상용 패키지인 edgeR package를 탑재하여 생산된 표적 RNA시퀀싱 데이터로부터 자동적으로 정규화된 유전자 발현정보를 추출하는 파이프라인을 설계할 수 있다.
- [121] NGS 기술을 이용해 생성된 시퀀싱 데이터는 사용 정렬 소프트웨어(예컨대, RNA-STAR)를 이용하여 참조 유전체에 맵핑된다. 맵핑 결과를 통해 각각의 유전자로부터 나온 서열의 개수를 집계할 수 있고, 이는 유전자의 발현량에 대한 직접적인 추정치가 된다.
- [122] 정규화 파이프라인은 맵핑이 완료되어 BAM 파일 형식으로 가공된 데이터를 입력받는다. 맵핑된 데이터는 파이프라인에 내장된 일련의 소프트웨어 패키지들 HTseq-count 및 edgeR 에 의해 샘플간 비교가 가능한 정규화된 발현량 값으로 계산될 수 있다.
- [123] 도 6은 RNA 시퀀스 데이터를 정규화하는 과정에 대한 예이다. 도 6의 두 개의 서로 다른 샘플에 대한 데이터를 정규화하는 과정의 예이다. 먼저 샘플 데이터는 유전자 발현량의 의미하는 매핑된 데이터를 입력받는다. 컴퓨터 장치는 HTseq-count를 적용하여 유전자 발현량을 계산한다. 각각의 샘플은 라이브러리 크기가 다르다(샘플 1은 100서열, 샘플 2는 300서열). 이후 컴퓨터 장치는 edgeR을 적용하여 발현량을 정규화한다. 도 6은 라이브러리 크기 100을 기준으로 샘플의 발현량을 정규화한 예이다.
- [124] 컴퓨터 장치는 유전자 발현량 데이터를 사전에 마련한 기계학습모델에 입력한다(140). 기계학습모델에 입력되는 유전자 발현량 데이터는 정규화된 데이터이다. 기계학습모델은 사전에 타겟 유전자의 발현량과 Oncotype DX RS로 학습된 것이다. 예컨대, 기계학습모델은 학습과정에서 Oncotype DX RS가 25이상인 경우 고위험(재발 가능성 높음)이라고 판단하고, 25미만인 경우 저위험이라고 판단할 수 있다.
- [125] 도 7은 기계학습의 인공신경망 구조에 대한 예이다. 도 7은 기계학습 모델 중 인공신경망(Artificial Neural Network)에 대한 예이다. 인공신경망은 다양한 종류 중 어느 하나를 활용할 수 있다. 도 7의 인공신경망은 완전 연결 구조(fully connected network)이다. 단계별로 은닉 노드(hidden node)를 사용한다. 또한

학습데이터에 대한 오버피팅(over-fitting)을 막기 위하여 가가 계층별로 배치 정규화(batch normalization)를 적용할 수 있다. 배치 정규화는 계층을 통과한 결과에 대하여 활성화값을 적절하게 강제로 분포시키는 과정이다.

- [126] 기계학습모델에서 설명한 바와 같이 인공신경망도 복수의 샘플에 대한 데이터를 이용하여 사전에 학습된다. 즉 복수의 샘플에 대한 타겟 유전자 발현량을 입력값을 이용하고, 해당 타겟 유전자(샘플)의 Oncotype DX RS를 이용하여 인공신경망 모델을 생성한다.
- [127] 이제 특정 피험자의 데이터를 신경망에 입력하여 해당 피험자의 유방암 예후를 예측할 수 있다. 사전에 학습된 인공신경망에 입력되는 값은 피험자의 타겟 유전자의 발현량이다. 인공신경망의 출력값은 Oncotype DX RS에 기반한 유방암 예후 예측 결과일 수도 있다. 인공신경망의 출력값은 특정한 점수와 같은 정량적 결과가 출력될 수 있다. 또는 인공신경망의 출력값은 곧바로 고위험 또는 저위험과 같은 정성적 평가 결과일 수도 있다.
- [128] 도 7에서 인공신경망을 중심으로 설명하였으나, 다양한 다른 기계학습모델을 이용하여 유방암 예후를 추정할 수 있다. 예컨대, 기계학습모델은 유전 알고리즘, 서포트 벡터 머신, 베이스 네트워크 등과 같은 다양한 방식을 활용할 수도 있다.
- [129] 컴퓨터 장치는 기계학습모델의 결과값을 기준으로 현재 입력된 샘플(피험자)에 대한 유방암 예후를 예측한다(150). 예컨대, 컴퓨터 장치는 기계학습모델의 결과값이 특정한 점수일 수 있다. 이 경우 컴퓨터 장치는 인공신경망의 출력값을 사전에 마련된 기준값과 비교하여 출력값이 특정값 이상인 경우, 현재 샘플(피험자)에 대하여 고위험이라고 판단할 수 있다. 반대로 컴퓨터 장치는 인공신경망의 출력값이 기준값 미만인 경우 저위험이라고 판단할 수 있다.
- [130] 도 8은 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후예측 시스템에 대한 예이다. 도 8(A)는 네트워크에서 구현된 시스템(200)의 예이다. 유방암 예후 예측 시스템(200)은 클라이언트 장치(210) 및 분석 서버(220)를 포함한다. 나아가 유방암 예후 예측 시스템(200)은 모델 DB(230)를 포함할 수도 있다. 분석 서버(220)는 전술한 컴퓨터 장치에 해당한다.
- [131] 클라이언트 장치(210)는 피험자에 대한 데이터를 제공하는 장치이다. 클라이언트 장치(210)는 피험자의 유전자 발현에 대한 데이터(유전자 데이터)를 분석 서버(220)에 전달한다. 클라이언트 장치(210)가 전달하는 데이터는 유전자 서열을 나타내는 데이터일 수 있다.
- [132] 분석 서버(220)는 유전자 데이터에 대한 NGS(next-generation sequencing) 기반 시퀀싱을 수행한 RNA 시퀀싱 데이터를 이용하여 타겟 유전자의 발현량을 측정하고, 사전에 마련한 기계학습모델에 상기 타겟 유전자의 발현량을 입력하여 출력되는 출력값을 기준으로 상기 피험자에 대한 유방암 예후를 추정한다. 분석 서버(220)는 전술한 데이터 후처리 및 정규화 과정을 수행할 수도

있다. 분석 서버(220)는 사전에 마련한 기계학습모델을 이용한다.

기계학습모델은 별도의 모델 DB(230)에 저장될 수도 있다. 분석 서버(220)가 데이터를 처리하고, 유방암 예후를 예측하는 과정은 전술한 바와 같다.

- [133] 도 8(B)는 유방암 예후 예측하는 컴퓨터 장치(300)에 대한 예이다. 도 8(B)에 도시한 컴퓨터 장치(300)는 전술한 분석 서버(220)일 수도 있다. 컴퓨터 장치(300)는 PC, 노트북, 스마트기기 또는 서버 등과 같은 장치를 의미한다. 컴퓨터 장치(300)는 입력장치(310), 연산장치(320), 저장장치(330) 및 출력장치(340)를 포함한다.
- [134] 입력장치(310)는 피험자의 유전자 데이터를 입력받는다. 유전자 데이터는 타겟 유전자의 발현에 관련된 데이터 내지 유전자 서열을 의미한다. 입력장치(310)는 피험자의 데이터를 통신이나 별도의 저장 장치를 통해 컴퓨터 장치(300)에 입력하는 장치이다. 나아가 입력장치(310)는 컴퓨터 장치(300)를 통해 피험자의 임상 데이터를 직접 입력받는 인터페이스 장치(키보드, 마우스, 터치 스크린 등)일 수도 있다.
- [135] 저장장치(330)는 전술한 기계학습모델을 저장하는 장치이다. 저장장치(330)는 입력장치(310)로부터 전달받은 피험자의 유전자 데이터를 저장할 수 있다. 저장장치(330)는 유전자 데이터를 분석하여 유방암 예후를 예측하는 소프트웨어가 설치된 장치일 수도 있다.
- [136] 연산 장치(320)는 입력된 유전자 데이터에 대한 유전자 발현량을 측정하고, 유전자 발현량을 기계학습모델에 입력하고, 기계학습모델의 출력값을 기준으로 피험자의 유방암 예후를 추정한다.
- [137] 출력장치(340)는 예후에 대한 정보를 일정한 형태로 출력하는 장치이다. 출력장치(340)는 디스플레이 장치, 문서를 출력하는 출력 장치 및 예후에 대한 정보를 다른 장치에 전달하는 통신 장치 중 적어도 하나를 포함한다.
- [138] 또한, 상술한 바와 같은 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법은 컴퓨터에서 실행될 수 있는 실행 가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수 있다. 상기 프로그램은 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.
- [139] 비일시적 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.
- [140] 본 실시예 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는

변형 예와 구체적인 실시례는 모두 전술한 기술의 권리범위에 포함되는 것이
자명하다고 할 것이다.

청구범위

- [청구항 1] 컴퓨터 장치가 피험자 조직(tissue)의 RNA 시퀀싱 데이터를 이용하여 타겟 유전자의 발현량을 측정하는 단계;
상기 컴퓨터 장치가 상기 타겟 유전자의 발현량을 사전에 마련한 인공신경망(Artificial Neural Network)에 입력하는 단계; 및
상기 컴퓨터 장치가 상기 인공신경망의 출력값을 기준으로 상기 피험자에 대한 유방암 예후를 추정하는 단계를 포함하되,
상기 인공신경망은 복수의 샘플의 타겟 유전자 발현량을 입력값으로 갖고, 상기 복수의 샘플에 대한 온코타입(Oncotype DX)의 재발 점수(recurrence score)에 기반한 결과를 출력하도록 사전에 마련되는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법.
- [청구항 2] 제1항에 있어서,
상기 컴퓨터 장치는 NGS(next-generation sequencing) 기법으로 상기 타겟 유전자의 발현량을 측정하는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법.
- [청구항 3] 제1항에 있어서,
상기 타겟 유전자는 온코타입 재발 점수 연산에 사용되는 16개의 유전자 및 상기 재발 점수와 상관도가 기준값 이상인 추가 유전자를 포함하되,
상기 추가 유전자는 ER(Estrogen) 수용체가 양성되면서 임파선 전이가 없는 샘플의 유전자들 각각에 대하여 상기 재발 점수와 상관 분석(Correlation Analysis)을 수행하여 상관 계수가 기준값 이상인 유전자들로 선택하는 차세대 염기서열분석을 이용한 기계 학습 기반 유방암 예후 예측 방법.
- [청구항 4] 제1항에 있어서,
상기 RNA 시퀀싱 데이터는 복수의 샘플 데이터에서 공통되게 발현되는 보존된 엑손(conserved exon) 영역을 포함하는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법.
- [청구항 5] 제1항에 있어서,
상기 컴퓨터 장치는 TMM(Trimmed Mean of M-value) 기법을 이용하여 상기 타겟 유전자의 발현량을 정규화하고, 정규화된 발현량을 상기 인공신경망에 입력하는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법.
- [청구항 6] 제1항에 있어서,
상기 인공신경망은 완전 연결 구조(fully connected)인 3개의 은닉 계층을 포함하고, 각 계층별로 배치 정규화(batch normalization)가 적용되는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법.

- [청구항 7] 제1항에 있어서,
상기 컴퓨터 장치는 상기 출력값이 기준값 이상인 경우 상기 피험자가 고위험이라고 예후를 추정하고, 상기 출력값이 기준값 미만인 경우 상기 피험자가 저위험이라고 예후를 추정하는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법.
- [청구항 8] 컴퓨터에서 상기 제1항 내지 제7항 중 어느 하나의 항에 기재된 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 방법을 실행하기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체.
- [청구항 9] 피험자 조직(tissue)의 RNA 유전자 데이터를 저장하는 클라이언트 장치; 및
상기 유전자 데이터에 대한 NGS(next-generation sequencing) 기반 시퀀싱을 수행한 RNA 시퀀싱 데이터를 이용하여 타겟 유전자의 발현량을 측정하고, 사전에 마련한 인공신경망(neural network)에 상기 타겟 유전자의 발현량을 입력하여 출력되는 출력값을 기준으로 상기 피험자에 대한 유방암 예후를 추정하는 분석 서버를 포함하되,
상기 인공신경망은 복수의 샘플의 타겟 유전자 발현량을 입력값으로 갖고, 상기 복수의 샘플에 대한 온코타입(Oncotype DX)의 재발 점수(recurrence score)에 기반한 결과를 출력하도록 사전에 마련되는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 시스템.
- [청구항 10] 제9항에 있어서,
상기 타겟 유전자는 ER(Estrogen) 수용체가 양성인면서 임파선 전이가 없는 샘플의 유전자들 각각에 대하여 상기 재발 점수와 상관 분석(Correlation Analysis)을 수행하여 상관 계수가 기준값 이상인 유전자들을 포함하는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 시스템.
- [청구항 11] 제9항에 있어서,
상기 RNA 시퀀싱 데이터는 복수의 샘플 데이터에서 공통되게 발현되는 보존된 엑손(conserved exon) 영역을 포함하는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 시스템.
- [청구항 12] 제9항에 있어서,
상기 분석 서버는 TMM(Trimmed Mean of M-value) 기법을 이용하여 상기 타겟 유전자의 발현량을 정규화하고, 정규화된 발현량을 상기 인공신경망에 입력하는 차세대 염기서열분석을 이용한 기계학습 기반 유방암 예후 예측 시스템.