한국인 고유유전체 참조표준

Korean Reference Genome Construction

류제운, 김대수, 박종화

한국생명공학연구원 국가생물자원정보관리센터

ABSTRACT

한국인 최초 전체 유전체 서열(KOREF; Korean individual genome sequence) 은 한국인을 위한 참조 서열로 써 사용될 수 있다. 2009년 1월에 남성 한국인 유전체를 솔렉사(Solexa)를 통해 전장서열을 결정하였다. 이는 NCBI의 인간게놈프로젝트에서 생산한 게놈의 99.83%를 커버하며, 또한 NCBI게놈서열의 약 20배를 커버할 정도의 유전체 서열을 결정하여 매우 높은 정확도를 가진 한국인 고유유전체이다. 한국인 유전체 서열의 분석결과 현재까지 밝혀지지 않았던 한국인 특이적인 3백만 개의 SNP를 밝혀냈다. 먼저 보고된 중국인 게놈은 한국인 게놈과 매우 가까운 민족 그룹임에도 불구하고 38% (3,186,352 SNP중에 1,217,362 SNP) 의 특이적인 차이를 나타내었으며, 또한 미토콘드리아 서열 비교를 통해서도 특이적인 다양성을 보여주는 SNP데이터를 확인 할 수 있었다. 차세대 게놈서열결정의 기술은 적은 노력과 비용으로 인간 유전체 데이터를 얻을 수 있게 되었으며, 이러한 개인유전체 데이터는 개인유전체 의학으로 가는 초석이 될 것이다.

Keyword: 개인유전체, 한국인 고유 유전체)

1. 서론

1977 년 Sanger [1]에 의해 처음으로 바이러스 전체 게놈 서열을 밝혀졌으며, 그로부터 3 년 후에는 Andreson [2]에 의해 휴먼 미토콘드리아 유전체가 밝혀졌다. 그 후 2004 년 처음으로 인간 유전체 서열결정이 완료된 이후 [3] 최신 분자생물학적인 실험기법의 눈부신 발달로 인해 [4]유전체 서열결정의 놀랄만한 기술의 진보를 이루었다.

Sanger dideoxy method 에 의해 최초 개인 유전체가 결정되었지만, NGS(next generation sequencing) 방법은 돈과 노력이 극적으로 줄어들기 때문에 개인 유전체에 매우 적합하다 [4]. 최근 NGS 에 의해 분석된 개인 유전체

서열이 발표되었다 [5,6,7]. 엄밀히 말하면, NGS 는 NCBI 의 인간 게놈프로젝트에서 만들어낸 인간 유전체를 바탕으로 인간 유전체를 재 서열 결정하는데 유효하다. 값비싸고 노력이 많이 드는 de novo assembly 에 기반한 유전체서열을 만드는 데신 이미 알려진 인간 유전체 전장서열에 NGS 유전체 서열결정방법에 의해 생산된 매우 짧은 read 를 맵핑은 적은 노력과 시간을 드려 분석이 가능하다. NGS 가 소개된 이후로, 전체 유전체 서열을 결정하는데 있어서 데이터 이동, 저장, 정확한 서열정렬, 다양한 질병과의 연관분석 등이 수행되는데 필요한 다양한 최신 생명정보학적 분석방법의 개발과 분석파이프라인들이 개발이 절실히 필요하게 되었다.

NCBI 인간게놈 프로젝트에서 완성한 인간 게놈 유전체는 NGS 로 유전체서열 재결정 필수적이지만 염기 배열과 유전자 수를 결정적인 측면에서는 이상적인 그림을 따르지 않는다. 게다가 최근 연구에 따르면 Wang [6]은 염기 read 의 13%가 NCBI 유전체 서열에 서열정렬이 되지 않음을 확인할 수 있었다. 이는 민족간의 거리가 멀기 때문에 유전체 서열이 정렬 되지 않은 것이다. 다른 말로 표현하면 소수민족을 대표할 수 있는 유전체 서열 재결정은 민족적으로 서로 다른 유전체들을 분석하기 위해서는 꼭 필요하다 할 수 있다. 이것은 한국인 게놈 시컨싱을 시작한 여러 이유 중 하나가 될 수 있다. 한국인과 중국인은 수 천년 동안 같은 조상으로부터 유래된 것으로 생각된다. 백인 게놈과 관련돼 있는 이 두 게놈의 다양성을 비교해보는 것은 우리 서로가 어떻게 다른지에 대한 통찰력을 제공해주고 있다.

이에 우리는 처음으로 전체 길이의 한국인 개인 전장유전체(KOREF)를 보고한다. KOREF 는 한국인 참조 게놈을 위한 국책 사업의 일환으로 미래의 한국인 개인 게놈 프로젝트를 위한 참조 기준으로 쓰여질 것이다.

2. Materials and Methods

DNA 추출. QIAamp DNA blood kit 의 사용설명서에 따라 혈액에서 Genomic DNA (gDNA)를 뽑아낸다.

Short read alignment. 빠른 짧은-리드 정렬 프로그램인 MAQ(ver. 0.7)을 사용하였다. MAQ 는 짝을 이루는 끝 쪽 read 의 read-pair 정보를 활용한다. 만약 mate pair 가 잘 정렬되어있다면, MAQ 는 이런 read-pair 정보를 이용하여 잘못된 정렬을 고치고 맞는 정렬을 추가하고, 정확하게 반복적인 유전체 서열을 맵핑한다..

Calling SNPs

MAQ 를 이용하여 짧은 read 를 NCBI 참조 게놈에 서열정렬하고, 서열정열 된 결과로부터 일치된 유전자형 서열을 생산하였다. SNP 로 부르기 위해서는 다음의 조건을 만족시켜야 한다. 최소 read depth 는 4 이며 (-d 4), 무작위로 위치한 개별적인 hit 을 제거하기 위한 최대 depth 는 100 (-D 100), consensus 품질은 20 (Q20), 인접한 서열 품질은 20 (Q20) 이여야 한다. 그리고 만약 3bp 의 flanking region 에서 어떤 indel 이 나타난다면 그것은 SNP 라고 하지 않는다.

3. Results and Discussions

우리는 50 기가바이트의 서열(36 베이스 read 를 통한 약 1,071,000,000 pair, 75 베이스 read 를 통한 약 154,000,000 pair)을 얻었다. MAQ (Mapping and Assembly with Qualities) [8] 프로그램을 이용하여 서열의 44Gb (89.67% of all data)는 NCBI 인간 게놈 reference (build 36.1)와 유전체 서열을 정렬 하였다. 전체적으로 NCBI reference 게놈의 99.83%가 평균적으로 15.5 fold mapping depth 를 통해 유전체 서열이 정렬되었다. KOREF 게놈의 SNP 를 HuRef(Venter's Genome)와 비교해보았다 (그림 1) [6,9]. KOREF 는 YH (Chinese Genome) 게놈에서 약 62%의 SNP 를 HuRef 게놈에서는 51%의 SNP 를

공유하는 한편 YH 게놈과 HuRef 게놈은 51%를 공유한다.

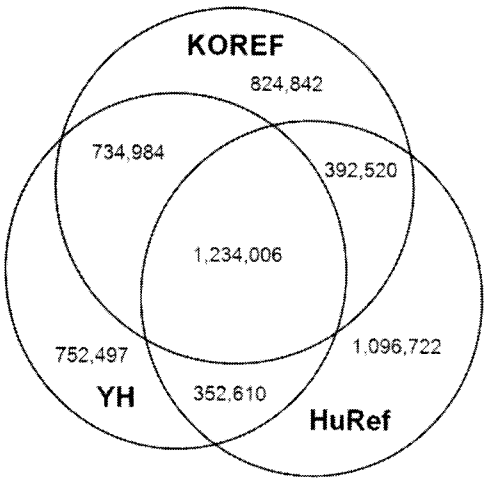

그림 1. Comparison of SNPs among KOREF, HuRef, and YH.

또한 MAQ 를 이용하여 229,187 개의 짧은 삽입과 결실 (-31 - +14 bp)를 확인하였다. 전체 229,187 개의 유전체 서열의 삽입과 결실 중에서 유전자영역에서 발생한 유전체 삽입과 결실을 분석해본, 유전자 부분에 위치한 85,393 indel 을 찾을 수 있었다. 유전자정보는 UCSC genome browser database (Kuhn et al. 2009)로부터 단백질로 발현되는 유전자 셋을 사용하여 KOREF 의 exon, CDS, UTRs, intron 에서의 indel 의 빈도수를 조사하였다. 229,187 개의 유전체 삽입과 결실 중에 37.25%에 해당하는 유전체서열의 삽입과 결실이 5' UTR (19 indel), 3' UTR (229 indel), CDS (38 indel), 그리고 intron (85,107 indel) 등 이미 알려져 있는 유전자에 위치하였다 (표 1). CDS 에서 찾은 38 개의 유전체 서열의 삽입과 결실 중에서 23 개의 유전자에서 발생한 유전체서열의 삽입과 결실이 단백질로 만들어지는 과정에 영향을 주어 단백질의 기능에 변화를 줄 수 있는 있다. 이와 같은 사실은 최근의 공통된 조상으로부터 분기된 이후 발생한 것으로서 NCBI 인간 유전체와 KOREF 유전체는 서로 다른 진화의 압력을 받고 있다 할 수 있다.

표 1. Indels in korean reference genome

Index	Indel			
	Indel number	Homozygous	Heterozygous	Gene number
5'UTR	19	7	12	32
CDS	38	12	26	27
3'UTR	229	81	148	162
Intron	85,107	32,150	52,957	10,317
Total	85,393	32,250	53,143	10,538

최근 연구에서 우리가 찾은 3,100,000 SNP 중에서 36,000 SNP (11.4%) 이상이 새로운 SNP 이라고 밝혔다. 비록 Wang (Wang et al. 2008)이 최근에 처음으로 중국인 중에서 한족을 대상으로 유전체 서열을 결정 하였지만, SNP 관련 연구들을 보면 중국인, 일본인 그리고 한국인을 포함한 아시아 소수 그룹이 최근에 유전적으로 갈려졌음을 알 수 있다 [10,11,12]. 그림 1 을 보면, KOREF 는 HuRef 와 같은 백인의 게놈에 비해 YH 와 좀더 많은 SNP 를 공유함을 알 수 있다. 그럼에도 불구하고, KOREF 와 YH 는 여전히 1,200,000 SNP (약 36%) 상당의 유전적 차이를 가지고 있다. 그리고 KOREF 와 YH 간의 유전체 서열의 삽입과 결실을 비교하면 두 게놈 사이에 약 11%정도만의 유전체 서열의 삽입과 결실을 공유하며, KOREF 와 HuRef 는 2.9%의 유전체서열의 삽입과 결실을 공유한다.

아마 KOREF 와 YH 간의 유전체서열의 정렬이 되지 않은 서열을 비교하면 새로운 민족의 특이적인 유전적 유전체 대한 새로운 통찰력을

얻을지도 모른다. 요컨대 이러한 연구결과는 가장 가까운 민족적 그룹간에서의 유전적 차이는 상당히 중요하다는 것을 나타낸다. 특히 하나의 소수 민족 그룹을 대표할 수 있는 민족의 유전체서열을 결정하면 전반적인 인종간의 유전적 차이를 정확하고 효율적으로 분석할 수 있다.

References

[1] Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265: 687-695.

[2] Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F. et al. 1981. Sequence and organization of the human mitochondrial genome. Nature 290: 457-465.

[3] International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.

[4] Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. Nat Biotechnol 26: 1135-1145.

[5] Bentley, D.R. Balasubramanian, S. Swerdlow, H.P. Smith, G.P. Milton, J. Brown, C.G. Hall, K.P. Evers, D.J. Barnes, C.L. Bignell, H.R. et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59.

[6] Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y. et al. 2008. The diploid genome sequence of an Asian individual. Nature 456: 60-65.

[7] Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-876.

[8] Li, H., Ruan, J., and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858.

[9] Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. et al. 2007. The diploid genome sequence of an individual human. PLoS Biol 5: e254.

[10] Hammer, M.F., Karafet, T.M., Park, H., Omoto, K., Harihara, S., Stoneking, M., and Horai, S. 2006. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet 51: 47-58.

[11] Jin, H.J., Tyler-Smith, C., and Kim, W. 2009. The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers. PLoS ONE 4: e4210.

[12] Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R.S., Redd, A.J., Zegura, S.L., and Hammer, M.F. 2001. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet 69: 615-628.