보도자료

미래를 개척하는 지식 공동체
서 울 대 학 교
SEOUL NATIONAL UNIVERSITY

보도일시	제한없음(즉시) / 2022. 10.7.(금)
문의	담당자: 고병준(02-880-4822)
	연구단장/연구책임자 김희발 교수(02-880-4803) / 교신저자

농생대 김희발 교수팀, 기존 척추동물 유전체에서 수천개의 유전자 조립 오류 발견

- 유전체 분야 권위지 Genome Biology에 논문 2편 동시 게재 -

■ 요약

연구 필요성	**- 서울대학교 김희발 교수팀은 2018년 공식 출범한 척추동물 유전체 프로젝트(Vertebrate genome project, 이하 'VGP')의 리더인 에릭자비스 교수(록펠러 대학)와 척추동물의 표준유전체 구축과 관련하여 긴밀한 국제 공동연구를 수행해 왔다.** **- 표준유전체 자료는 향후 모든 생물학 연구의 기반이 되므로 표준유전체는 그 자체로 무결해야 한다는 것을 연구팀은 관련 학계에 강조해왔으나, 표준유전체 전반에 포함된 오류의 양과 유형을 계측한 연구는 찾아보기 힘들었다.**
연구성과/ 기대효과	**- 서울대학교 김희발 교수팀은 VGP와의 협업 수행을 통해 기존에 조립된 표준유전체로부터 수천개의 유전자 조립 오류를 발견하여 유전체 분야 권위지인 지놈바이올로지(Impact factor 17.91)에 이례적으로 관련 논문 2편을 동시 온라인 게재(9월 27일) 하였다.** **- 이 연구성과는 향후 국내외 표준유전체 구축의 방향성을 제시한다는 점에서 중요하며 이를 바탕으로 국내 교수팀이 국제 표준유전체 구축 분야를 선도할 수 있는 계기를 마련했다는 점에서 의미하는 바가 크다.**

■ 본문

□ 서울대학교 김희발 교수팀은 국제 공동연구인 척추동물 유전체 프로젝트(Vertebrate genome project, 이하 VGP)와의 협업 수행을 통해 기존에 조립된 표준유전체로부터 수천개의 유전자 조립 오류를 발견하여 **유전체 분야 권위지인 지놈바이올로지(Impact factor 17.91)에 이례적으**

로 관련 논문 2편을 동시 온라인 게재(9월 27일) 하였다.

□ VGP는 지구상에 서식하는 66,000여 종의 척추동물 표준유전체 구축 및 유용 유전자 발굴을 목적으로 하는 전 세계적 프로젝트로 인간과 동물의 질병 연구는 물론 생물의 진화와 멸종위기종 보전 연구에 적용 가능한 유전체를 구축하는 것이 목적이다. 서울대학교 김희발 교수팀은 2018년 공식 출범한 VGP의 리더인 에릭자비스 교수(록펠러 대학)와 척추동물의 표준유전체 구축과 관련하여 긴밀한 국제 공동연구를 수행해 왔다. 특히 이들은 **과거에 구축된 표준유전체와 새롭게 구축된 VGP 그룹의 표준유전체를 정량적으로 비교하는 체계를 구축**하였으며 이를 통해 **기존의 표준유전체가 포함하는 오류의 양과 유형을 정밀하게 계측**하였다. 표준유전체 자료는 향후 모든 생물학 연구의 기반이 되므로 표준유전체는 그 자체로 무결해야 한다는 것을 연구팀은 관련 학계에 강조해왔다. 이번에 새롭게 구축한 표준유전체 비교체계를 이용하여 생물학 분야의 모델 생물로 잘 알려진 **금화조, 벌새, 오리너구리, 등목어의 기존 표준유전체에서 각 종별 수백에서 수천개에 달하는 유전자의 결손 및 복제가 유전체 구축 과정 중에 발생한 오류였다는 사실을 발견**하였다. 이러한 결과는 기존의 과학적 통념을 넘어서는 많은 양의 유전자 조립 오류를 포함하는 것이다. 만약 이렇게 많은 양의 오류를 포함한 유전체를 기반으로 연구를 수행할 경우 연구자들이 잘못된 결론에 도달할 것은 자명한 것이다. 실제 이번 연구를 통해 과거 학계에 알려진 몇몇 금화조, 벌새, 오리너구리의 종 특이적 진화 특성 조차 유전체 조립과정에서 발생한 오류에 기인하였다는 것을 밝혀내었다.

□ 향후 진화학, 의학, 뇌과학 분야 등 생물학의 많은 부분이 유전체 빅데이터를 바탕으로 연구되어질 것이며 이들 연구의 기반이 되는 표준유전체의 중요성은 점점 부각될 것이다. 따라서 기존의 표준유전체에 만연해 있는 오류를 선제적으로 발견한 연구 성과는 향후 국내외 표준유전체 구축의 방향성을 제시한다는 점에서 중요하며 이를 바탕으로 국내 교수팀이 국제 표준유전체 구축 분야를 선도할 수 있는 계기를 마련했다는 점에서 의미하는 바가 크다.

□ **연구결과**

Ko et al. Genome Biology (2022) 23:205
https://doi.org/10.1186/s13059-022-02764-1

Genome Biology

RESEARCH Open Access

Widespread false gene gains caused by duplication errors in genome assemblies

Byung June Ko[1†], Chul Lee[2†], Juwan Kim[2], Arang Rhie[3], Dong Ahn Yoo[2], Kerstin Howe[4], Jonathan Wood[4], Seoae Cho[5], Samara Brown[6,7], Giulio Formenti[6], Erich D. Jarvis[6,7*] and Heebal Kim[1,2,5*]

†Byung June Ko and Chul Lee contributed equally.

*Correspondence: ejarvis@mail.rockefeller.edu; heebal@snu.ac.kr

[1] Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea
[2] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[3] Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, USA
[4] Wellcome Sanger Institute, Cambridge, UK
[5] eGnome, Inc, Seoul, Republic of Korea
[6] Laboratory of the Neurogenetics of Language, The Rockefeller University, New York, NY, USA
[7] Howard Hughes Medical Institute, Chevy Chase, MD, USA

Abstract

Background: False duplications in genome assemblies lead to false biological conclusions. We quantified false duplications in popularly used previous genome assemblies for platypus, zebra finch, and Anna's Hummingbird, and their new counterparts of the same species generated by the Vertebrate Genomes Project, of which the Vertebrate Genomes Project pipeline attempted to eliminate false duplications through haplotype phasing and purging. These assemblies are among the first generated by the Vertebrate Genomes Project where there was a prior chromosomal level reference assembly to compare with.

Results: Whole genome alignments revealed that 4 to 16% of the sequences are falsely duplicated in the previous assemblies, impacting hundreds to thousands of genes. These lead to overestimated gene family expansions. The main source of the false duplications is heterotype duplications, where the haplotype sequences were relatively more divergent than other parts of the genome leading the assembly algorithms to classify them as separate genes or genomic regions. A minor source is sequencing errors. Ancient ATP nucleotide binding gene families have a higher prevalence of false duplications compared to other gene families. Although present in a smaller proportion, we observe false duplications remaining in the Vertebrate Genomes Project assemblies that can be identified and purged.

Conclusions: This study highlights the need for more advanced assembly methods that better separate haplotypes and sequence errors, and the need for cautious analyses on gene gains.

Keywords: False duplication, Assembly error, Phasing error, De novo assembly, Vertebrate genome project

<Widespread false gene gains caused by duplication errors in genome assemblies>

기존에 구축된 금화조, 벌새, 오리너구리의 표준유전체 내 수백에서 수천개 유전자의 복제가 유전체 조립오류에 의해 발생한 것을 확인하였다. 이들 유전자 복제 오류의 양과 유형을 계측하였으며 향후 유전자 복제 오류를 방지하기 위해 유전체 조립 시 반수체 구분 및 염기서열 오류 방지의 중요성을 강조하였다.

Kim et al. Genome Biology (2022) 23:204
https://doi.org/10.1186/s13059-022-02765-0

Genome Biology

RESEARCH Open Access

False gene and chromosome losses in genome assemblies caused by GC content variation and repeats

Juwan Kim[1†], Chul Lee[1†], Byung June Ko[2], Dong Ahn Yoo[1], Sohyoung Won[1], Adam M. Phillippy[3], Olivier Fedrigo[4], Guojie Zhang[5,6,7,8], Kerstin Howe[9], Jonathan Wood[9], Richard Durbin[9,10], Giulio Formenti[4,11], Samara Brown[11], Lindsey Cantin[11], Claudio V. Mello[12], Seoae Cho[13], Arang Rhie[3], Heebal Kim[1,2,13*] and Erich D. Jarvis[4,11,14*]

†Juwan Kim and Chul Lee contributed equally to this work.

*Correspondence: heebal@snu.ac.kr; ejarvis@rockefeller.edu

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[4] Vertebrate Genome Lab, The Rockefeller University, New York City, USA
Full list of author information is available at the end of the article

Abstract

Background: Many short-read genome assemblies have been found to be incomplete and contain mis-assemblies. The Vertebrate Genomes Project has been producing new reference genome assemblies with an emphasis on being as complete and error-free as possible, which requires utilizing long reads, long-range scaffolding data, new assembly algorithms, and manual curation. A more thorough evaluation of the recent references relative to prior assemblies can provide a detailed overview of the types and magnitude of improvements.

Results: Here we evaluate new vertebrate genome references relative to the previous assemblies for the same species and, in two cases, the same individuals, including a mammal (platypus), two birds (zebra finch, Anna's hummingbird), and a fish (climbing perch). We find that up to 11% of genomic sequence is entirely missing in the previous assemblies. In the Vertebrate Genomes Project zebra finch assembly, we identify eight new GC- and repeat-rich micro-chromosomes with high gene density. The impact of missing sequences is biased towards GC-rich 5′-proximal promoters and 5′ exon regions of protein-coding genes and long non-coding RNAs. Between 26 and 60% of genes include structural or sequence errors that could lead to misunderstanding of their function when using the previous genome assemblies.

Conclusions: Our findings reveal novel regulatory landscapes and protein coding sequences that have been greatly underestimated in previous assemblies and are now present in the Vertebrate Genomes Project reference genomes.

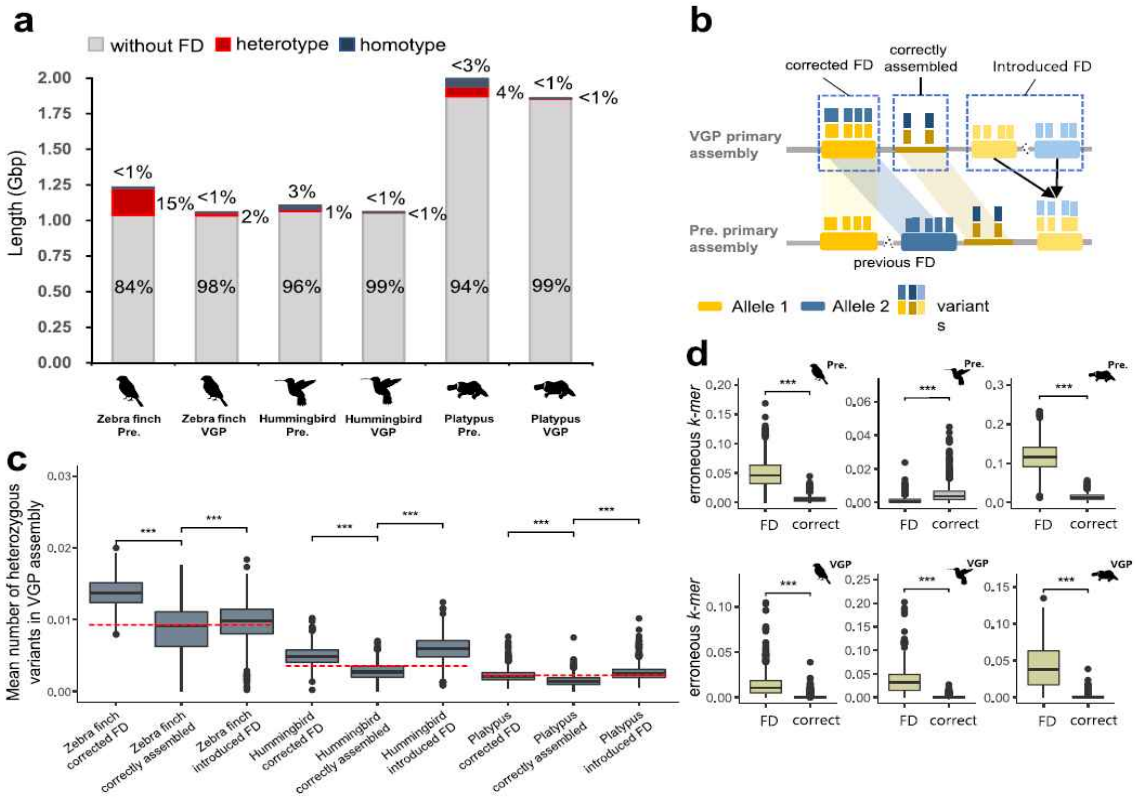
Keywords: Genomics, Gene structure, GC content, Genomic dark matter, Annotation

<False gene and chromosome losses in genome assemblies caused by GC content variation and repeats>

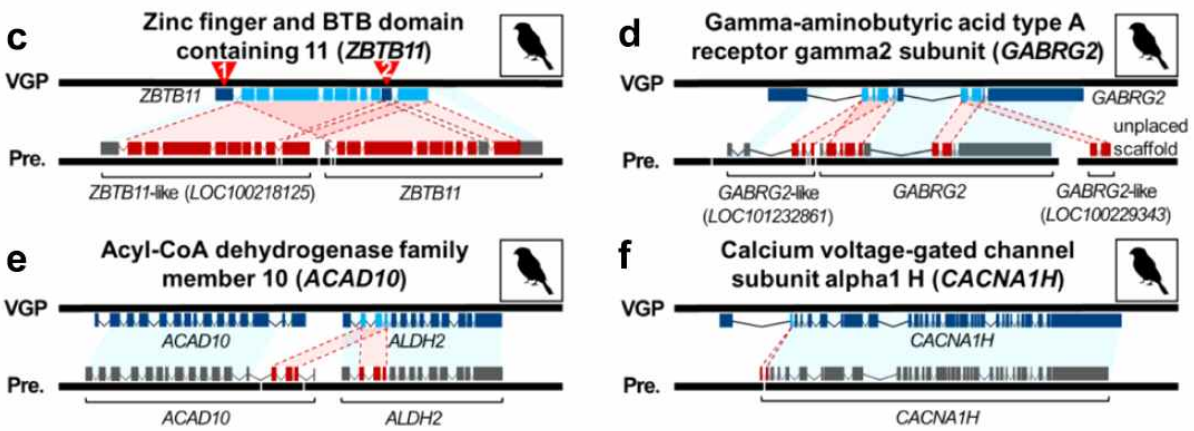
기존에 구축된 금화조, 벌새, 오리너구리, 등목어의 표준유전체 내 수천개의 유전자가 조립오류로부터 영향을 받아 전체 또는 일부가 손실된 것으로 나타났다. 특히 전체 유전자중 절반 이상의 유전자 프로모터 지역에서 손실오류가 발견되었으며 이는 높은 GC content에 의한 염기서열 해

독과정의 편향으로 부터 발생한 것으로 나타났다. 이 결과는 유전체의 조립과정에서 발생한 오류가 잘못된 생물학적 해석을 야기 할 수 있다는 것을 시사한다.

□ **그림설명**

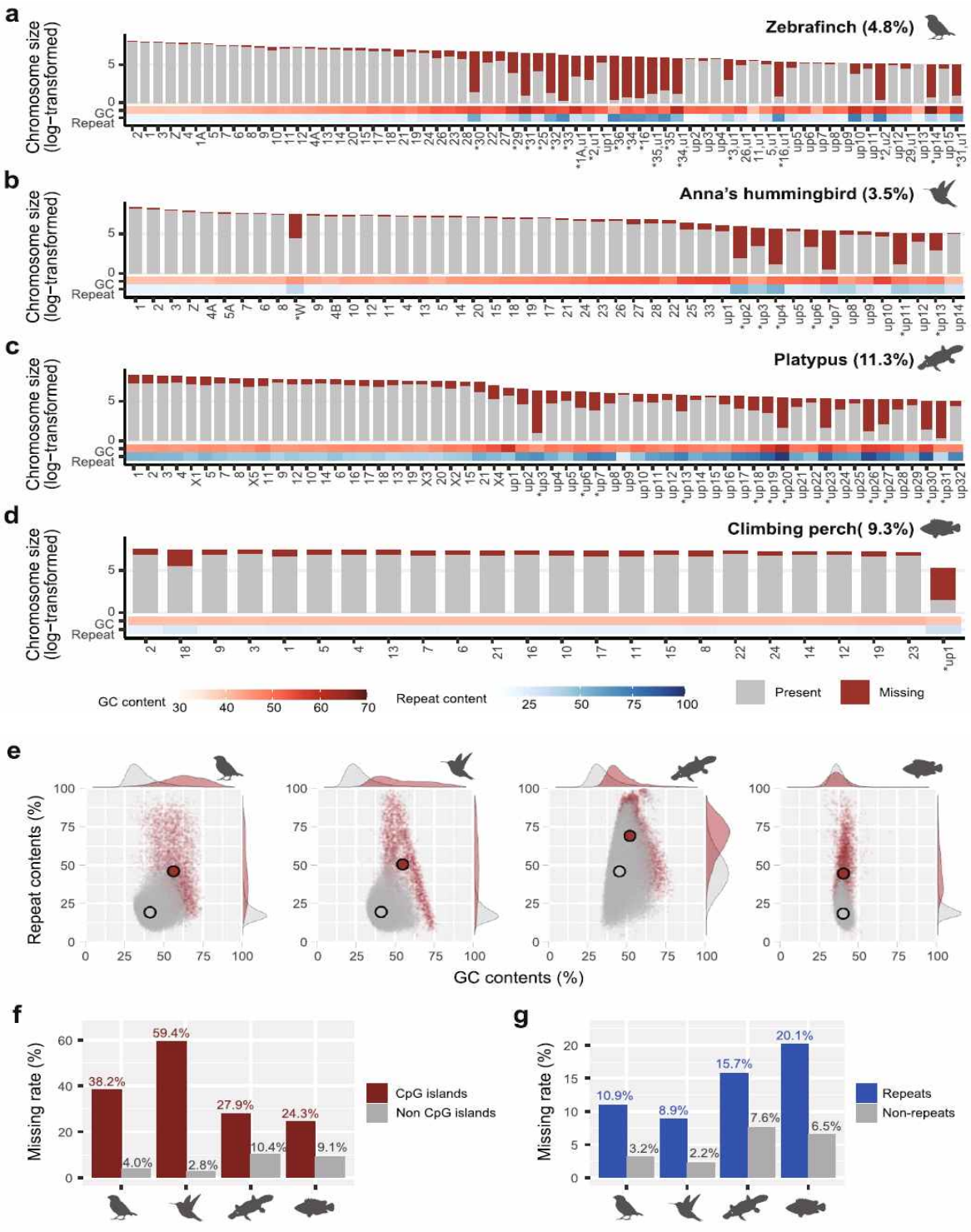


금화조, 벌새, 오리너구리의 유전자 조립 오류에 의한 허위 염기서열 복제 양을 계측하였으며, 이들 복제오류가 높은 이형접합성 및 염기서열 해독 오류와 관련이 있다는 것을 나타낸다.



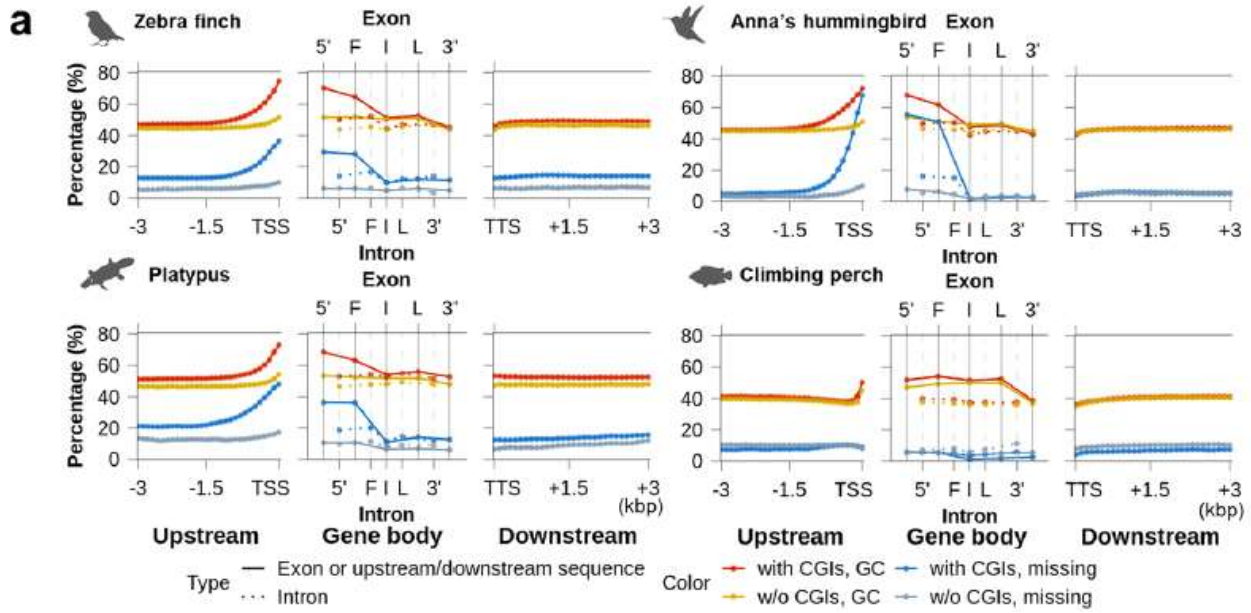
금화조 내 허위 유전자 획득의 네가지 사례. 위 사례는 유전자 전체 혹은 일부가 복제 오류에 의

해 새로운 유전자로 잘못 인식된 경우, 각기 다른 유전자가 하나의 유전자에 혼합된 경우, 하나의 유전자 내 엑손 영역이 허위 복제된 경우를 나타낸다.

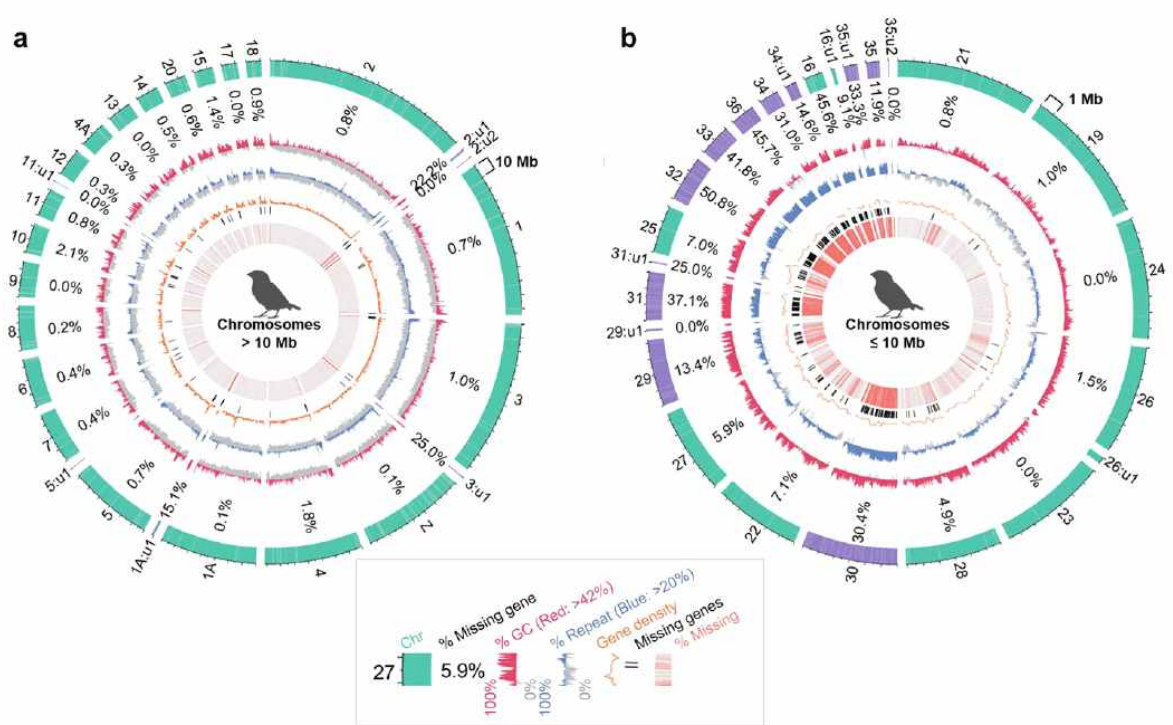


금화조, 벌새, 오리너구리, 등목어의 유전자 조립 오류에 의한 허위 유전자 손실 양을 계측하였으

며, 이들 손실 오류가 GC 및 repeat content와 연관이 있다는 것을 나타낸다.



금화조, 벌새, 오리너구리, 등목어의 표준유전체 유전자들의 5말단 지역에서 허위 손실 오류가 빈번히 발생한 것을 밝혀냈다.



금화조 표준유전체 전반에 걸쳐 허위 유전자 손실 오류를 계측하였으며 이를 통해 금화조의

micro chromosome들을 새롭게 발견하였다.

□ **연구자**

○ 성 명 : 김희발

○ 소 속 : 서울대학교 농생명공학부 교수

○ 연락처 : 02-880-4803, heebal@snu.ac.kr