De novo rates and selection of large copy number variation

Priit Palta

Bioinfo Journal Club

10.01.2011

Research

De novo rates and selection of large copy number variation

Andy Itsara,[1] Hao Wu,[2] Joshua D. Smith,[1] Deborah A. Nickerson,[1] Isabelle Romieu,[3,5] Stephanie J. London,[2] and Evan E. Eichler[1,4,6]

[1]*Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA;* [2]*National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina 27709 USA;* [3]*National Institute of Public Health, Cuernavaca, Morelos 62100, Mexico;* [4]*Howard Hughes Medical Institute, Seattle, Washington 98195, USA*

While copy number variation (CNV) is an active area of research, de novo mutation rates within human populations are not well characterized. By focusing on large (>100 kbp) events, we estimate the rate of de novo CNV formation in humans by analyzing 4394 transmissions from human pedigrees with and without neurocognitive disease. We show that a significant limitation in directly measuring genome-wide CNV mutation is accessing DNA derived from primary tissues as opposed to cell lines. We conservatively estimated the genome-wide CNV mutation rate using single nucleotide polymorphism (SNP) microarrays to analyze whole-blood derived DNA from asthmatic trios, a collection in which we observed no elevation in the prevalence of large CNVs. At a resolution of ~30 kb, nine de novo CNVs were observed from 772 transmissions, corresponding to a mutation rate of $\mu = 1.2 \times 10^{-2}$ CNVs per genome per transmission ($\mu = 6.5 \times 10^{-3}$ for CNVs >500 kb). Combined with previous estimates of CNV prevalence and assuming a model of mutation-selection balance, we estimate significant purifying selection for large (>500 kb) events at the genome-wide level to be $s = 0.16$. Supporting this, we identify de novo CNVs in 717 multiplex autism pedigrees from the AGRE collection and observe a fourfold enrichment ($P = 1.4 \times 10^{-3}$) for de novo CNVs in cases of multiplex autism versus unaffected siblings, suggesting that many de novo CNV mutations contribute a subtle, but significant risk for autism. We observe no parental bias in the origin or transmission of CNVs among any of the cohorts studied.

Genome Research, published online September 14, 2010

Problem

- While copy number variation (CNV) is an active area of research, **de novo mutation rates within human populations** are not well characterized

Data

Table 1. Overview of data sets

Study	Description	Source DNA	Array platform	No. of trios (before QC)[a]
Asthma	Trios, Mexico City; child with asthma	Blood	Illumina 550K	386 (492)
HapMap	Trios, Ibadan, Nigeria, and Utah; no ascertained phenotype	Cell line	Illumina 1M Duo	54 (59)
AGRE	Pedigrees, various locations; ≥1 case of autism or similar	Cell line	Illumina 550K	1757 (1996)
	Multiplex autism			1638
	Simplex autism			119

[a]The relatively large number of excluded trios is due to higher stringency for CNV discovery

HapMap – discovery of CNVs

- 54 trios
- 1366 CNVs called in kids that were manually screened
- No significant difference in the number of CNVs per individual between parents and children

HapMap – discovery of CNVs

- Among the inherited events that could be assigned to a single parent, there was no significant difference between maternal and paternal inheritance (374 vs. 333, binomial test P = 0.1324)

- After several additional filters, we identified **seven** candidate de novo CNVs ranging in size from 25-260 kb

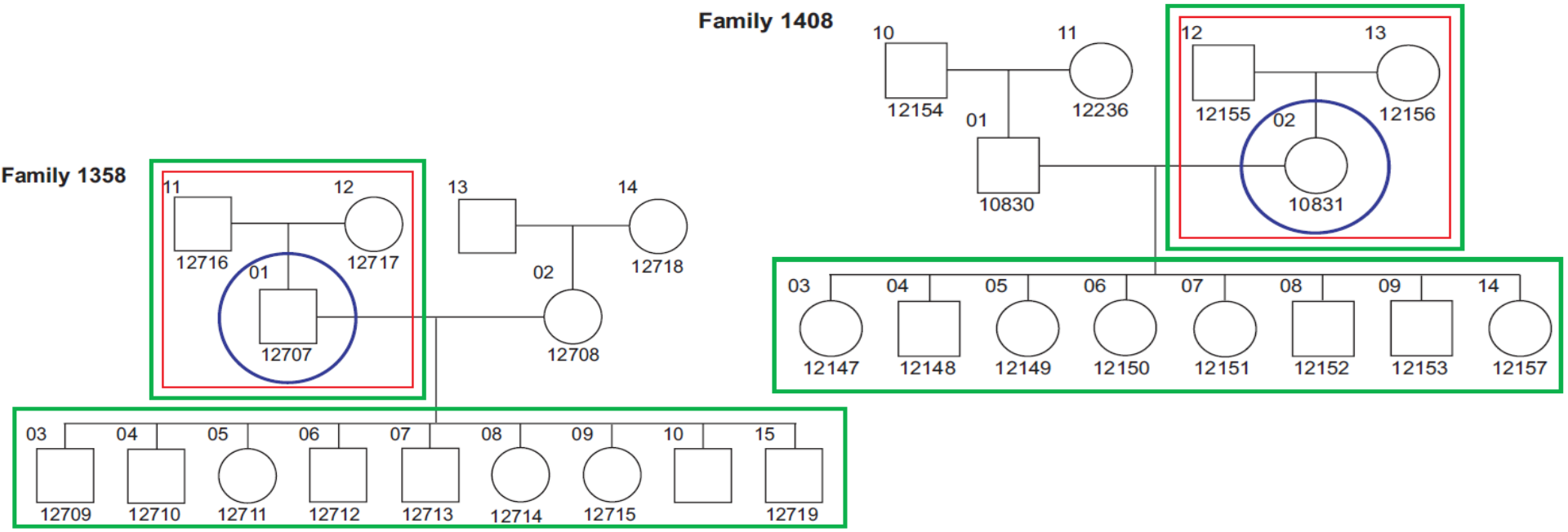
MapMap – validation of de novo CNVs (theory)

- A truly de novo CNV would be unobserved in the first generation (CEU trio parents), validated in the second generation (CEU trio children), and, assuming no selective effects, transmitted to approximately half of the individuals in the third generation

- Observing transmission of a CNV would serve to distinguish between a true CNV within the germline versus a potential cell line artifact

MapMap – validation of de novo CNVs

- While all **four** (tested) CNVs were validated by array-CGH in the second generation, trans-mission of these CNVs was never observed in the third generation (23? grandchildren)



MapMap – conclusion

- Putative de novo CNVs found in the HapMap data set likely represent cell line artifacts that arose during passaging of cell cultures or represent potential somatic mosaicisms that were cloned during the establishment of the cell culture

Asthma – discovery of CNVs

- 386 trios
- 2025 CNVs identified in children
- No significant difference in the number of CNVs per individual between parents and children

Asthma – discovery of CNVs

- Among inherited events, there were no significant differences in the paternally versus maternally inherited CNVs (490 vs. 522, binomial test P = 0.3298)

- Among the children, authors identified 11 (out of 2025) CNVs for which a corresponding CNV was not observed in either parent

Asthma – validation

- Validation was carried out by using custom array-CGH platform for **nine** candidate de novo CNVs
- For all available parental DNA, no CNVs were detected at these loci
- In the children, **eight** of nine loci validated as copy number changes
- One CNV was removed from further analysis based on its overlap with a known site of copy number polymorphism (CNP)

Asthma – conclusion

- Frequency and size distribution of CNVs was not significantly different from previous frequency estimates in the general population

- This implies that asthmatics are unlikely to be enriched in large CNVs and that this cohort allows estimates of de novo rates of CNVs that are applicable to the general population

Asthma – conclusion

- Authors estimate the genome-wide frequency of de novo CNVs to be 1.2 × 10-2 (9 out of 772) per haploid genome per generation

- Although this estimate is not significantly different from previous estimates, it is likely to be conservative, as it does not account for CNVs overlapping CNPs, regions for which there is inadequate probe coverage, and regions, such as segmental duplications (SDs), that are often refractory to CNV detection using array-based techniques

AGRE – discovery of CNVs

- Largest collections of autism families – Autism Genetics Resource Exchange
- 1757 trios
- 10 839 CNV calls in children

AGRE – discovery of CNVs

- No significant difference in the number of CNVs per individual between parents and children

- Again, no significant difference in the rate of maternal versus paternal inheritance of CNVs (3103 vs. 3059, binomial test P = 0.5838)

AGRE – de novo CNVs

- 67 putative de novo CNVs (62 in multiplex trios, five in simplex) representing 64 independent events
- Authors did observe a dramatic difference in rate of de novo mutation between affected and unaffected siblings from the same autism families
- Several de novo CNVs occurred both at loci previously associated with variable phenotypes, including autism as well as sites of recurrent CNV not previously associated with autism

AGRE – de novo CNVs

- Over one-quarter (17/64) of de novo CNVs in the AGRE collection were flanked by SDs in direct orientation (12/64) or had one of the breakpoints mapping within a cluster of SDs

AGRE - conclusion

- A fourfold enrichment ($P = 1.4 \times 10^{-3}$) for de novo CNVs in cases of multiplex autism versus unaffected siblings, suggesting that many de novo CNV mutations contribute a subtle, but significant risk for autism

Parental origin of de novo CNVs

- By analyzing the B-allele frequency (BAF) in de novo CNVs, authors were able to unambiguously determine the parental origin for 47 of 73 de novo CNVs in the asthma and AGRE data sets

- They identified 21 paternal and 26 maternal de novo events (mean size = 1.25 Mb and 1.19 Mb, respectively) and therefore no evidence for a parent-of-origin preference

Conclusions

- Significant limitation in directly measuring genome-wide CNV mutation is accessing DNA derived from primary tissues (e.g. blood) as opposed to cell lines

- No parental bias in the origin or transmission of CNVs among any of the cohorts studied

Conclusions

- Estimated genome-wide CNV mutation rate using single nucleotide polymorphism (SNP) microarrays to analyze whole-blood derived DNA: $\mu = 1.2 \times 10^{-2}$ CNVs per haploid genome per transmission ($\mu = 6.5 \times 10^{-3}$ for CNVs >500 kb)

Conclusions

- This estimate has the benefit of being a direct estimate based on approximately twice the number of trios as in previous works

- Although there is less uncertainty in our estimate of the genome-wide CNV mutation rate, it is likely to be conservative as it does not account for regions of the genome not adequately covered by our platforms, such as SDs or common sites of CNPs as these regions were excluded in our analysis. Conversely, SDs and CNPs cover ~5%–6%, so that our estimate is largely applicable to ~94% of the human genome