A roadmap for understanding the evolutionary significance of structural genomic variation

Claire Mérot[1*#], Rebekah A. Oomen[2,3*#], Anna Tigano[4,5*#], Maren Wellenreuther[6,7*#]

* all authors contributed equally

Corresponding authors claire.merot@gmail.com, rebekahoomen@gmail.com, anna.tigano@unh.edu, Maren.Wellenreuther@plantandfood.co.nz

[1]Université Laval, Institut de Biologie Intégrative des Systèmes, 1030 Avenue de la Médecine, G1V 0A6, Québec, QC, Canada

[2]Centre for Ecological and Evolutionary Synthesis, University of Oslo, Blindernveien 31, 0371 Oslo, Norway

[3]Centre for Coastal Research, University of Agder, Universitetsveien 25, 4630 Kristiansand, Norway

[4]Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH, USA

[5]Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH, USA

[6]School of Biological Sciences, The University of Auckland, Auckland, New Zealand

[7]The New Zealand Institute for Plant & Food Research Ltd, Nelson, New Zealand

KEYWORDS:

Adaptation - Chromosomal rearrangements - Copy number variants - Duplications - Inversions - Population genomics - Speciation

ABSTRACT

Structural genomic variants (SVs) are ubiquitous and play a major role in adaptation and speciation. Yet, comparative and, later, population genomics have focused predominantly on gene duplications and large-effect inversions. The lack of a common framework for studying all SVs is hampering progress towards a more systematic assessment of their evolutionary significance. Here we 1) review how different types of SVs affect ecological and evolutionary processes, 2) suggest unifying definitions and recommendations for future studies, and 3) provide a roadmap for the integration of SVs with eco-evolutionary studies. In doing so, we lay the foundation for population genomics, theoretical, and experimental approaches to understand how the full spectrum of SVs impacts ecological and evolutionary processes.

Beyond SNPs: structural variation plays a key role in adaptive evolution and speciation

The study of ***structural variants*** (***SVs***) (see Glossary, Figure 1) has a long history going back to the discovery of ***chromosomal inversions*** in *Drosophila* fruit flies in the early 20th century [1], followed by ***transposable elements*** *(**TEs**)* in maize (*Zea mays*) [2], and ***gene duplications*** in *Drosophila* [3]. Yet, this rich knowledge from comparative genetics was not widely integrated into the field of molecular population genetics, which rose in the 1970s. Since then, predominant attention has been on molecular markers that quantify patterns defined by one or few base pairs, such as ***SNPs***, ***AFLPs***, and ***microsatellites***. However, diverse forms of SVs have reemerged in population-level studies owing to advances in genomic technologies. Mounting evidence suggests that they are taxonomically ubiquitous [4–7] and key contributors to a multitude of evolutionary processes (Box 1; [8]).

Considering the full spectrum of structural variants

Large inversions — spanning 100 kb to several Mb — are the most frequent SVs associated with adaptive phenotypes and the maintenance of differentiation [9,10]. The strong association is largely due to their ease of detection and their ability to reduce recombination in inversion heterozygotes (***heterokaryotypes***), and hence to preserve linkage between alleles despite gene flow. Although they have received less attention, other SVs such as chromosomal ***fusions*** and ***translocations*** also interfere with recombination and promote differentiation. For example, a chromosomal fusion polymorphism in some Atlantic salmon (*Salmo salar)* populations in Canada is associated with precipitation and harbors five times stronger differentiation than neutral SNP variation [11]. The fusion of several chromosomes in *Heliconius* butterflies is associated with a higher speciation rate [12]. Indeed, karyotype engineering shows that chromosome fusions lead to the rapid emergence of reproductive isolation in *Saccharomyces cerevisiae* yeast [13]. Translocations can also be involved in speciation: in the house mouse *Mus musculus,* four incipient species with different karyotypes coexist in the Swiss-Italian Alps [14].

Gene duplication, and the subsequent evolution of novel functions, is probably the best documented effect of ***Copy Number Variants*** *(**CNVs**)* on adaptation and

diversification [15]. However, CNVs encompass a much wider class of variants, including ***insertions/deletions*** (***indels***), tandem repeats (***mini- and microsatellites***), and variation in copy number for a given coding or non-coding sequence. They represent the most common SV type and can modify gene dosage and reshape gene structure [16]. A large CNV linked to plumage dimorphism and thermal adaptation in common murres (*Uria lomvia*) appears to suppress recombination locally [17]. Copy number variation associated with toxin resistance has also been demonstrated multiple times, indicating that CNVs may enable rapid adaptation to environmental stressors [18]. Micro- and minisatellite data, used predominantly as neutral markers in the past (but see [19]), also represent a common type of SV with demonstrated functional impact [20,21].

Transposable elements (***TEs***) are major modifiers of genome structure [22] and drivers of adaptation and reproductive isolation [23]. TEs represent a type of translocation and/or duplication and a source of indels because they 'jump' from one location to another. TE insertions also lead to segmental duplications and inversions, due to ***non-allelic homologous recombination*** *[24]*. TEs can change during an individual's lifetime, which makes them an important variant in rapidly changing environments [25].

A better understanding of how structural variants affect evolutionary processes is needed

While recent studies provide exciting insights into the role of SVs in adaptation and diversification, they also reveal limitations that hamper progress. For example, many studies investigating the genomic basis of traits from sequence data have found a link between a phenotype and a SV, most often a large inversion or gene duplication (e.g., [18,26–28]. Whether such examples are representative of the global importance of SVs or if their prevalence is biased by their relative ease of detection is still unclear. However, with ever-improving sequencing and analytical methods, we can now adopt a bottom-up approach and explore genomes independently from phenotypes to identify SVs of different types and sizes that could be associated with different evolutionary processes. Generally, synthesis in the field is slowed by a lack of unified definitions and the absence of a framework to synthesize information from SVs and SNPs in population genomics. We suggest definitions and focus points to guide future investigations and propose a roadmap to integrate SVs into evolutionary genomics (Figure 2).

Defining and detecting structural variants of all types and sizes

Sequence and structural variation exist along a continuous spectrum

Definitions of biological phenomena reflect the thoughts and methods in the field that coined them. 'Chromosomal rearrangement' was used to describe inversions, fusions, and translocations detected at a microscopic scale using cytogenetics. The term 'structural variation' emerged in 2004 with its characterization in the human genome [29] and now generally refers to smaller-scale variants detected from sequence data. However, sequence and structural variation exists on a size spectrum ranging from ***Single Nucleotide Variants (SNVs)***, including SNPs and single nucleotide indels, up to large SVs affecting hundreds of Mb (Figure 1).

SVs are also classified according to how they alter the genome, i.e. whether they add, delete, or change the position or orientation of DNA (Figure 1). As highlighted by recent reviews on inversions [9,10,30], most studies focus on only one type of SV rather than considering their diversity. For example, CNVs and TEs are often not considered 'chromosomal rearrangements', resulting in an oversight of similarities shared among SVs. We argue that the field would benefit from jointly considering the full diversity of SVs and advocate for a wider adoption of the term 'structural variant' to encompass all changes in position or direction, as well as gains or losses of sequence, without imposing a size limit, to enable synthesis across studies.

Systematic characterization of structural variants of all types and sizes is needed

Regions of elevated differentiation linked to phenotypic variation and exhibiting signatures of ***linkage disequilibrium (LD)*** (Box 2) are often ascribed to inversions. However, such ***blocks of differentiation***, or ***haploblocks***, can likewise result from other types of SVs (e.g., CNVs [17], fusions [11]) or be due to selective sweeps [31] or introgression [32]. Follow-up analyses are needed to definitively associate a haploblock with a SV. Moreover, indirect identifications are biased towards large SVs (> 1 Mb) with large phenotypic effect and/or high sequence divergence, and overlook small, neutral, and recently established SVs.

Recent developments in sequencing and computational methods have enabled direct genome-wide characterization of SVs, providing information on SV position, frequency, breakpoints, and gene content [33,34] (Box 2). However, challenges remain. High-quality, chromosome-level reference genomes are seldom available, yet are helpful to localize and characterize SVs. Sampling enough individuals to capture the geographic, phenotypic, and sexual population variation is needed to characterize structural diversity [35], but can be logistically and financially prohibitive. Further, the sensitivity of different detection methods varies with respect to SV size [7,33] and is not generally reported. To enable comparisons and syntheses and identify best practices (e.g., data type, software, settings), we need simulations and benchmarking to test how detection power varies by analytical approach, SV type, and type of sequence data (Figure 2).

A framework for understanding the evolutionary significance of structural variation

Structural variants are missing pieces to the puzzle of genomic variation

SVs might explain some of the 'missing heritability' in many genotype-phenotype association studies [36]. In the crow *Corvus corone*, a retrotransposon indel of 2.25 kb explained an additional 10% in plumage colouration variance between two subspecies compared to SNP variation only [7]. ***eQTL*** studies that integrate CNVs and SNPs in humans identified several SVs that cause gene expression changes, often with larger effect sizes than SNPs [37,38]. Signatures of population structure can also vary depending on the type of marker. In modern humans, CNVs and deletions show different signatures of population structure and selection, with the former revealing a stronger spatial signature [39]. Moreover, SVs can encompass two to five times more bases of the genome than SNPs [4,40]. SVs also follow different evolutionary trajectories. For instance, some large inversions are under long-term balancing selection and are involved in interspecific introgression [41], while TEs and microsatellites commonly evolve rapidly [21,25]. Therefore, accounting for the range of genetic variation requires going beyond SNPs and integrating SVs into studies investigating genome evolution, levels of standing genetic variation, population structure, demography, phenotype-genotype associations, and the genomic basis of adaptation and speciation.

Population genomics can reveal the roles of SVs in evolutionary processes

Cost-effective ways to analyse SVs at larger scales in non-model species are emerging. For example, CNVs and large inversions can be genotyped, directly or indirectly (Box 2), using low-coverage whole-genome sequencing [42] or RAD-seq [27,43,44]. Complex and large SVs are better characterized by long-range information (Box 2), but these methods can be expensive. New tools are necessary to leverage information from a subset of diverse and well-sequenced genomes to genotype SVs in larger datasets.

Some analytical methods developed for traditional markers may be used to mine information on SVs from existing population-scale datasets. For instance, population genomics based on CNVs uses an extension of the F_{ST} index of differentiation called V_{ST} [45]. Coding SVs similarly to SNPs and genotyping different SVs for large numbers of individuals is a challenge. CNVs can be relatively easily summarized in a matrix of read depths, but expressing genotypes as numbers of copies remains difficult. For balanced SVs (Figure 1), analyses can either focus on SNPs genotyped within the rearranged region [5] or consider the SV as an individual locus, with the latter being a more powerful approach to finding associations with phenotypic and environmental variation [46].

The joint analysis of SNPs and SVs in a population genomics framework will allow us to test whether sequence differentiation associated with SVs has adaptive value or is due to demographic and population structure (e.g. [44]). Systematic analysis of SVs will address the detection bias towards large inversions and help to unveil how different features of SVs (e.g., size, position, content, type, breakpoints) influence evolutionary trajectories (e.g., [47]). Comparing SNPs and different kinds of SVs will reveal factors causing variability in evolutionary rates across the genome. Finally, comparing numbers and distributions of SVs among populations connected by varying levels of gene flow will improve our understanding of how gene flow-selection balance affects the genomic architecture of adaptive traits [48]. Altogether, such studies will enable us to shed light on when and how SVs form, persist, and spread among populations and species (e.g., *de novo* formation or introgression, drift, balancing or fluctuating selection).

Theoretical approaches are needed to predict evolutionary patterns specific to SVs

Theoretical models have been pivotal to developing hypotheses on why SVs might follow a different evolutionary pathway compared to SNPs [49–51]. Models have shed light on TE dynamics [52] and the role of recombination suppression in adaptation with gene flow, particularly in inversions [49–51]. Less is known about the evolutionary significance of other features of SVs, such as the multi-allelic characteristics of CNVs, the impacts of reduced effective population sizes (N_e) of inversions and deletions, and differences in mutation rates within SVs. Theoretical studies targeting a wider variety of SVs are needed to understand how different features relate to their origin and maintenance, and the relative contribution of selective and neutral processes in their evolution.

Forward individual-based simulations are a promising tool to account for SV complexity under realistic evolutionary scenarios. For instance, the program *SLiM 3* [53] models population genetic processes and includes genetic variation based on SNPs and TEs, and information on LD. Such simulations enable evaluating the relative effects of migration, drift, and selection on SV dynamics (e.g., [54]) and, reciprocally, to predict the conditions under which SVs represent relevant architectures for adaptation and differentiation [51]. Forward simulations can model expected signatures of selective and demographic processes, enabling comparisons between simulated and empirical data to identify the specific processes and range of conditions that explain SV distributions in natural populations. Simulated genomic data are also useful for testing the performance of genome-scan methods [55], especially regarding the effects of SVs on detecting putative targets of selection [56].

Backward simulations based on coalescent theory can also contribute to our understanding of SV evolution. Comparing demographic models sheds light on the evolutionary history of SVs [41,57]. Such simulations enable comparisons of coalescence times across different parts of the genome, or between different variant types, populations, or species. They provide a projection of the expected polymorphism frequencies under neutrality, against which the distribution of SVs can be contrasted [58]. Thus, backward simulations are another way of disentangling the contributions of demographic and selective processes to creating observed SV frequencies.

Experiments can reveal the mechanisms by which SVs impact phenotypes

Common garden and reciprocal transplant experiments comparing groups with different SV genotypes are classic approaches for demonstrating adaptation [59,60]. However, care must be taken to account for differences in genomic background. Combining numerous artificial crosses with statistical modelling can help to separate the effects of SVs from the rest of the genome, yet genetically modifying SVs into alternate genomic backgrounds in a full factorial design would be ideal.

Experimental evolution approaches can test theoretical predictions about the genomic architecture of polygenic traits. This approach revealed alternate genomic architectures underlying the evolution of growth rate in the marine fish *Menidia menidia* following size-selective harvesting. An extended haploblock was implicated in the evolution of smaller sizes in one experimental population but not its replicate, where evolutionary changes were associated with unlinked SNPs [61].

Analyses of gene expression can shed further light on the adaptive roles of SVs and has supported the ***recombination suppression hypothesis*** [49,60,62] and direct gene effects near breakpoints [63] (Box 1). Strong support for the recombination suppression hypothesis was found in *D. melanogaster* by comparing gene expression patterns between natural inversions, which influenced expression genome-wide, and genetically engineered synthetic inversions, which had negligible effects on expression [64]. Gene expression analyses can reveal gene dosage effects of CNVs on associated phenotypes [16]. Experimental knockdown of genes inside rearrangements can be used to functionally annotate SVs [28].

There is a pressing need for experiments directed towards understanding the effects of SVs on recombination. High resolution sequencing of offspring, heterozygous for the SV of interest, can be used to measure recombination rates of regions within and proximal to SVs [65]. Note that the effects of recombination suppression can be diluted by ***gene conversion***, whose rates within SVs can be quantified using a similar approach [66].

Concluding remarks and future perspectives

The field of structural genomic variation has matured to move beyond the most easily detected variants and to investigate the mechanisms underlying the relevance of all SVs for evolution. As more high-quality genome assemblies become available, we expect SVs to be investigated in an increasing number and diversity of non-model organisms.

Future syntheses of these studies will provide new insights into several outstanding questions regarding the respective roles of structural and sequence variation in evolution, differences in abundances and distributions of SVs among taxa, how SVs relate to ecological specialization, and how they affect recombination. By cataloguing the whole spectrum of genetic variation, we will gain insights into the mechanisms that create genomic hotspots of diversity. Because evolutionary dynamics of SVs differ from other parts of the genome, they will help us tease apart evolutionary and demographic effects on genome evolution that were hitherto hidden. Resurrecting classic micro- and minisatellite data and treating them as SVs might facilitate a better understanding of the role of these variants in evolutionary processes (but see [67]). Further, systematic inclusion of SVs in both empirical and theoretical studies will enable a better understanding of the roles of selection, drift, and gene flow in SV maintenance and how population connectivity across large and small scales impacts SV distribution and evolution.

In the future, SVs will be integrated into ecological and evolutionary applications such as conservation genomics, plant and animal breeding, and global change biology, as well as applications based on ancient and environmental DNA. It is therefore fundamental that we enable future comparisons across studies and taxa by developing generalizable tools and best practices in order to maximize the ecological and evolutionary insights provided by the joint analysis of genome sequence and structure.

Acknowledgments

We are grateful to Louis Bernatchez for initially gathering us around this topic. We would like to thank H. T. Baalsrud, L. Bernatchez, Y. Dorant, N. Gemmell, S. Jentoft, M. Laporte, M. Newberry, C. Rougeux, the UNVEIL network and two anonymous reviewers for constructive discussions about SVs and comments on the manuscript. Special thanks to E.

Berdan, R. Buri, B. Fuller, K. Johannesson, D. Lowry, T. Nelson, A. Terceira, M. Weissensteiner, J. Wolf, and D. Lowry for providing photographs for Figure I (in Box 1), and J. Ditner for drawings of *Uria aalge*. We would like to acknowledge the funding agencies that have supported our work on SVs (CM: Fonds de Recherche du Québec & Banting fellowships; RAO: James S. McDonnell Foundation Postdoctoral Fellowship 220020556; AT: National Institutes of Health NIH R35GM128843; MW: MBIE Endeavour Research Programme: C11X1603, Royal Society of New Zealand Marsden MFP-PAF1801, Swedish Research Council Vetenskapsrådet 2012-3996).

REFERENCES

1 Sturtevant, A.H. (1913) The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J. Exp. Zool.* 14, 43–59

2 McClintock, B. (1931) *Cytological observations of deficiencies involving known genes, translocations and an inversion in Zea mays*, University of Missouri, College of Agriculture, Agricultural Experiment Station.

3 Bridges, C.B. (1936) The bar “gene” a duplication. *Science* 83, 210–211

4 Catanach, A. *et al.* (2019) The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. *Mol. Ecol.* 28, 1210–1223

5 Lucek, K. *et al.* (2019) The role of structural genomic variants in population differentiation and ecotype formation in *Timema cristinae* walking sticks. *Molecular Ecology* 28, 1224–1237

6 Chakraborty, M. *et al.* (2019) Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* 10, 4872

7 Weissensteiner, M.H. *et al.* (2019) The population genomics of structural variation in a songbird genus. *bioRxiv*, 830356

8 Wellenreuther, M. *et al.* (2019) Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.* 28, 1203–1209

9 Wellenreuther, M. and Bernatchez, L. (2018) Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* 33, 427–440

10 Faria, R. *et al.* (2019) Evolving inversions. *Trends Ecol. Evol.* 34, 239–248

11 Wellband, K. *et al.* (2019) Chromosomal fusion and life history-associated genomic variation contribute to within-river local adaptation of Atlantic salmon. *Molecular Ecology* 28, 1439–1459

12 Davey, J.W. *et al.* (2016) Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3* 6, 695–708

13 Luo, J. *et al.* (2018) Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature* 560, 392–396

14 Giménez, M.D. *et al.* (2017) A half-century of studies on a chromosomal hybrid zone of the house mouse. *J. Hered.* 108, 25–35

15 Ohno, S. (2013) *Evolution by Gene Duplication*, Springer Science & Business Media.

16 Gamazon, E.R. and Stranger, B.E. (2015) The impact of human copy number variation on gene expression. *Brief. Funct. Genomics* 14, 352–357

17 Tigano, A. *et al.* (2018) A complex copy number variant underlies differences in both colour

plumage and cold adaptation in a dimorphic seabird. *bioRxiv*, 507384
18 Reid, N.M. *et al.* (2016) The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* 354, 1305–1308
19 Buonaccorsi, V.P. *et al.* (2011) Characterization of a genomic divergence island between black-and-yellow and gopher Sebastes rockfishes. *Mol. Ecol.* 20, 2603–2618
20 Ranathunge, C. *et al.* (2018) Transcribed microsatellite allele lengths are often correlated with gene expression levels in natural sunflower populations. *bioRxiv*, 339903
21 Bagshaw, A.T.M. (2017) Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol. Evol.* 9, 2428–2443
22 Casacuberta, E. and González, J. (2013) The impact of transposable elements in environmental adaptation. *Mol. Ecol.* 22, 1503–1517
23 Serrato-Capuchina, A. and Matute, D.R. (2018) The role of transposable elements in speciation. *Genes* 9, 254
24 Klein, S.J. and O'Neill, R.J. (2018) Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res.* 26, 5–23
25 Rey, O. *et al.* (2016) Adaptation to global change: a transposable element–epigenetics perspective. *Trends Ecol. Evol.* 31, 514–526
26 Nelson, T.C. *et al.* (2019) Extreme copy number variation at a tRNA ligase gene affecting phenology and fitness in yellow monkeyflowers. *Mol. Ecol.* 28, 1460–1475
27 Huang, K. *et al.* (2019) Multiple chromosomal inversions contribute to adaptive divergence of a dune sunflower ecotype. *bioRxiv,* 829622
28 Nishikawa, H. *et al.* (2015) A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. *Nat. Genet.* 47, 405–409
29 Feuk, L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.* 7, 85–97
30 Hoffmann, A.A. and Rieseberg, L.H. (2008) Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* 39, 21–42
31 Chen, H. *et al.* (2010) Population differentiation as a test for selective sweeps. *Genome Res.* 20, 393–402
32 Duranton, M. *et al.* (2018) The origin and remolding of genomic islands of differentiation in the European sea bass. *Nat. Commun.* 9, 2518
33 Ho, S.S. *et al.* (2019) Structural variation in the sequencing era. *Nat. Rev. Genet.*1-19
34 Mahmoud, M. *et al.* (2019) Structural variant calling: the long and the short of it. *Genome Biology* 20, 246
35 Audano, P.A. *et al.* (2019) Characterizing the major structural variant alleles of the human genome. *Cell* 176, 663–675
36 Eichler, E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* 11, 446–450
37 Chiang, C. *et al.* (2017) The impact of structural variation on human gene expression. *Nat. Genet.* 49, 692–699
38 Hsieh, P. *et al.* (2019) Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* 366, eaax2083
39 Sudmant, P.H. *et al.* (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science* 349, aab3761
40 Tigano, A. *et al.* (2019) Comparative and population genomics approaches reveal the basis of adaptation to deserts in a small rodent. *bioRxiv*, 856310
41 Korunes, K.L. *et al.* (2019) Inversions shape the divergence of *Drosophila pseudoobscura* and *D. persimilis* on multiple timescales. *bioRxiv*, 842047
42 Hill, T. and Unckless, R.L. (2019) A deep learning approach for detecting copy number variation in next-generation sequencing data. *G3* 9, 3575–3582

43 McKinney, G.J. *et al.* (2017) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol. Ecol. Resour.* 17, 656–669

44 Dorant, Y. *et al.* (2020) Copy number variants outperform SNPs to reveal genotype-temperature association in a marine species. *bioRxiv*, 923490

45 Redon, R. *et al.* (2006) Global variation in copy number in the human genome. *Nature* 444, 444–454

46 Todesco, M. *et al.* (2019) Massive haplotypes underlie ecotypic differentiation in sunflowers. *bioRxiv*, 790279

47 McBroome, J. *et al.* (2019) Fine-scale position effects shape the distribution of inversion breakpoints in Drosophila melanogaster. *bioRxiv*, 793364

48 Tigano, A. and Friesen, V.L. (2016) Genomics of local adaptation with gene flow. *Mol. Ecol.* 25, 2144–2164

49 Kirkpatrick, M. and Barton, N. (2006) Chromosome inversions, local adaptation and speciation. *Genetics* 173, 419–434

50 Guerrero, R.F. and Kirkpatrick, M. (2014) Local adaptation and the evolution of chromosome fusions. *Evolution* 68, 2747–2756

51 Yeaman, S. (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl. Acad. Sci. U. S. A.* 110, E1743–51

52 Charlesworth, B. and Charlesworth, D. (1983) The population dynamics of transposable elements. *Genet. Res.* 42, 1–27

53 Haller, B.C. and Messer, P.W. (2019) SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Mol. Biol. Evol.* 36, 632–637

54 Berdan, E.L. *et al.* (2019) Muller's ratchet and the long-term fate of chromosomal inversions. *bioRxiv*, 606012

55 Lotterhos, K.E. and Whitlock, M.C. (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol. Ecol.* 23, 2178–2192

56 Lotterhos, K.E. (2019) The effect of neutral recombination variation on genome scans for election. *G3* 9, 1851–1867

57 Cayuela, H. *et al.* (2019) Standing genetic variation and chromosomal rearrangements facilitate local adaptation in a marine fish. *bioRxiv*, 782201

58 Xue, A.T. *et al.* (2018) Differential effect of selection against LINE retrotransposons among vertebrates inferred from whole-genome data and demographic modeling. *Genome Biol. Evol.* 10, 1265–1281

59 Wellenreuther, M. *et al.* (2017) Local adaptation along an environmental cline in a species with an inversion polymorphism. *J. Evol. Biol.* 30, 1068–1077

60 Cheng, C. *et al.* (2018) Systems genetic analysis of inversion polymorphisms in the malaria mosquito *Anopheles gambiae*. *Proc. Natl. Acad. Sci. U. S. A.* 115, E7005–E7014

61 Therkildsen, N.O. *et al.* (2019) Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science* 365, 487–490

62 Coughlan, J.M. and Willis, J.H. (2019) Dissecting the role of a large chromosomal inversion in life history divergence throughout the *Mimulus guttatus* species complex. *Mol. Ecol.* 28, 1343–1357

63 Huang, Y.-C. *et al.* (2018) Multiple large inversions and breakpoint rewiring of gene expression in the evolution of the fire ant social supergene. *Proc. Biol. Sci.* 285, 20180221

64 Said, I. *et al.* (2018) Linked genetic variation and not genome structure causes widespread differential expression associated with chromosomal inversions. *Proc. Natl. Acad. Sci. U. S. A.* 115, 5492–5497

65 Rowan, B.A. *et al.* (2019) An ultra high-density *Arabidopsis thaliana* crossover map that refines the influences of structural variation and epigenetic features. *Genetics* 213, 771–787

66 Korunes, K.L. and Noor, M.A.F. (2019) Pervasive gene conversion in chromosomal inversion heterozygotes. *Mol. Ecol.* 28, 1302–1315
67 Sunnucks, P. *et al.* (2000) SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. *Mol. Ecol.* 9, 1699–1710
68 Stewart, N.B. and Rogers, R.L. (2019) Chromosomal rearrangements as a source of new gene formation in *Drosophila yakuba*. *PLoS Genet.* 15, e1008314
69 Rieseberg, L.H. (2001) Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16, 351–358
70 Crown, K.N. *et al.* (2018) Local inversion heterozygosity alters recombination throughout the genome. *Curr. Biol.* 28, 2984–2990.e3
71 Elgin, S.C.R. and Reuter, G. (2013) Position-effect variegation, heterochromatin formation, and gene silencing in *Drosophila*. *Cold Spring Harb. Perspect. Biol.* 5, a017780
72 Paulsen, J. *et al.* (2019) Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nat. Genet.* 51, 835–843
73 Lavington, E. and Kern, A.D. (2017) The effect of common inversion polymorphisms In(2L)t and In(3R)Mo on patterns of transcriptional variation in *Drosophila melanogaster*. *G3* 7, 3659–3668
74 Mérot, C. *et al.* (2019) Balancing selection via life-history trade-offs maintains an inversion polymorphism in a seaweed fly. *Nat. Commun.* 11, 1-11
75 Jay, P. *et al.* (2019) Mutation accumulation in chromosomal inversions maintains wing pattern polymorphism in a butterfly. *bioRxiv*, 736504
76 Faria, R. *et al.* (2019) Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Mol. Ecol.* 28, 1375–1393
77 Oomen, R.A. *et al.* Consequences of single-locus and tightly linked genomic architectures for evolutionary responses to environmental change. *bioRxiv*, 928770
78 Li, H. and Ralph, P. (2019) Local PCA shows how the effect of population structure differs along the genome. *Genetics* 211, 289–304
79 Kosugi, S. *et al.* (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 20, 117
80 Van't Hof, A.E. *et al.* (2016) The industrial melanism mutation in British peppered moths is a transposable element. *Nature* 534, 102–105
81 Chaisson, M.J.P. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* 10, 1-16
82 Bianco, S. *et al.* (2018) Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* 50, 662–667
83 Chakraborty, M. *et al.* (2018) Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat. Genet.* 50, 20–25
84 Sherman, R.M. *et al.* (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* 51, 30–35
85 Paten, B. *et al.* (2017) Genome graphs and the evolution of genome inference. *Genome Res.* 27, 665–676
86 Gong, L. *et al.* (2018) Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat. Methods* 15, 455–460

FIGURES

Figure 1. The diversity of structural variants. Genetic variants vary in size from a single nucleotide to hundreds-of-Mb-long structural variants (SVs). SVs are classified according to how they change the genome sequence. Balanced SVs change the position and/or order of genomic areas. Unbalanced SVs involve a gain or loss of sequence. Note that transposable elements can cause translocations, indels, and/or duplications. SNV = Single Nucleotide Variant, including SNPs and single nucleotide indels; MNV = Multiple Nucleotide Variant; CNV = Copy Number Variant.

Figure 2 (Key Figure): A roadmap for understanding the evolutionary significance of structural genomic variation. Colors indicate different steps toward understanding the role of SVs in adaptation and speciation, from top to bottom.

Figure I (in Box 1). The effects of SVs on adaptation and speciation at multiple levels of biological organization. From bottom to top and left to right: CELL: Example mechanisms by which SVs impact the genome, from DNA sequence to chromosome. Effects of SVs on gene expression include changes in the distance between genes and their regulatory elements, chromatin state, and gene dosage. ORGANISM: Multiple copies of tRNA ligase in the yellow monkeyflower *Mimulus guttatus* are associated with shorter flowering time, leading to differential survival in dry years and variation in seed production [26] (photo by D. Lowry). A large CNV in the common murre *Uria aalge* is associated with differences in plumage and thermal adaptation [17] (drawings by J. Ditner). A 25 Mb inversion in the seaweed fly *Coelopa frigida* affects a life-history trade-off between larval survival and reproductive success [74] (photo by M. Wellenreuther). DIVERSIFICATION: The crab- and wave-ecotypes of *Littorina saxatilis* periwinkles harbour more than 17 chromosomal inversions whose frequencies vary between the two microhabitats despite gene flow, suggesting that they are involved in local adaptation [76] (photo by F. Pleijel). Two subspecies of European crow, *Corvus corvus corvis* and *C. corvus corone*, differ by a 2.25 kb retrotransposon insertion that affects plumage, a trait involved in pre-mating isolation [7] (photos by R. Burri). Genomic incompatibilities leading to reduced hybrid fitness and reproductive isolation between the bluefin (*Lucania goodei*) and rainwater (*L. parva*) killifish are associated with a Robertsonian fusion of the sex chromosome [54] (photos by A. Terceira).

Figure II (in Box 2). Overview of complementary approaches for SV detection

Sequencing: Reduced-representation sequencing (RRS) approaches target a fraction of the genome (e.g., RAD-seq and SNP-chips). WGS = whole-genome sequencing. CMS = connected molecule strategies. A chromosome-level genome assembly is usually necessary for the analyses of SVs (but see alternative approaches in [44,57]).
Indirect detection: “Local PCA” refers to principal component analyses performed on windows along the genome. The PCA in the haploblock region highlights a typical pattern with three clusters of individuals, corresponding to the three haploblock variant combinations [11,27,46,76,78]. In contrast, the PCA outside the haploblock shows no clustering.
Direct detection: SV detection algorithms are based on sequencing depth, read orientation, and read splitting of short and long reads [4,5,34,35]. RRS provides information on sequencing depth, enabling detection of CNVs [44,45]. Long-reads provide high resolution of SV breakpoints [86]. Hi-C links are chromatin contacts between pairs of loci represented by a triangular heatmap of the number of links. Accumulation of links between distant loci reveals SVs between the target sample and reference [46,81]. Linked-reads are short reads tagged with the same barcodes when originating from the same original DNA fragment (up to 100 kb). SVs can be detected from the long-range information carried by barcoded linked-reads [40]. The comparison of genetic maps [27,76], optical maps [7], or full assemblies [6,7] enables the detection of both intra- and inter-chromosomal rearrangements. We refer to “large SV” when >100 kb (Figure 1).

TEXT BOXES

Box 1: Structural variants affect the evolution and maintenance of adaptive traits and reproductive barriers at several levels of biological organization (Figure I).

At the genome level, structural variants (SVs) necessarily alter the linear structure (i.e., sequence) of DNA. These changes can affect the order and proximity of genetic elements and disrupt functionality of extant genes, or form new ones, by coupling or uncoupling promoters and coding regions [68]. Changes to DNA sequence can affect three-dimensional genome structure by altering folding patterns and histone interactions. SVs can form secondary structures during meiosis in heterozygotes that can interfere with recombination to varying degrees [65,69]. Suppression of recombination can occur through production of unbalanced meiotic products and by displacement of crossing-overs away from SVs [70]. Some SVs (e.g., fissions and fusions) change the number and size of chromosomes, thereby impacting recombination rates even within homokaryotypes.

SV impacts the transcriptome in several ways. An underappreciated mechanism, **Position-Effect Variegation** (**PEV**; [71]), occurs when changes in the spatial proximity of the DNA sequence to telomeres and centromeres, and thus heterochromatic regions, alters the expression levels of nearby genes. SVs can also change the proximity of regulatory elements to genes, potentially affecting gene expression across the genome [64]. Changes in the positions of genetic elements relative to histones and interactions among topologically associated domains can affect the exposure of transcription binding sites, thereby silencing or enhancing transcription [72]. Local effects of SVs on expression include changes in gene dosage [16], expression of *de novo* genes [68], loss of expression of genes disrupted by SV breakpoints or deletions, and alterations of the epigenetic environment near breakpoints [63,73]. If the SV is associated with reduced recombination, it can maintain LD among genes and regulatory elements [73].

SVs underlie diverse morphological, physiological, behavioural, and life history traits [8] and impact fitness through effects on survival and reproduction [74]. When SVs affect recombination, heterokaryotypes can experience partial sterility due to the formation of lethal or inviable recombinant products during meiosis [30]. A lack of recombination

prevents purging of deleterious mutations, resulting, over time, in higher fitness of heterokaryotypes [54,75].

SVs are frequently associated with various stages of diversification, including local adaptation [76], pre-mating isolation [7], and speciation [9,54]. Blocks of differentiation are predicted to be favoured under adaptation with gene flow [48] and are expected to alter the evolutionary trajectory of polygenic traits under selection as they resemble single loci of large effect, rather than many loci of small effect [77].

Box 2: Moving from indirect evidence to the direct detection of SVs (Figure II)

Indirect evidence: haploblocks of differentiation

An increasing number of studies are uncovering genetic differentiation driven by a subset of co-localized linked SNPs using unsupervised methods such as Principal Component Analysis (PCA) [76,78]. The combination of high differentiation and LD suggests that these SNPs may be associated with a SV reducing recombination. Based on this observation, sliding-window PCAs along the genome were employed to screen for these signatures across *Helianthus* sunflower ecotypes, which identified 37 haploblocks [46]. Similarly, inversions associated with two periwinkle (*Littorina saxatilis*) ecotypes were identified based on clusters of SNPs in LD [76]. Complementary evidence, including higher heterozygosity in putative heterokaryotypes, and recombination and heritability estimates based on genetic maps, can support the presence of an inversion [27].

Direct evidence: making the best of different sequencing methods to catalog SVs

Standard shotgun libraries (i.e., with short insert size, generally < 1 kb) sequenced with Illumina short reads are the most common type of sequencing data and can be used to directly detect SVs (reviewed in [79]). However, they are not necessarily the best for identifying SVs, particularly large ones. Mate-pair libraries have more power than shotgun libraries to detect SVs because their paired reads have larger insert sizes (> 1 kb) and are more likely to span SV boundaries [5]. Additionally, SVs are often associated with repeats and duplications that are difficult to assemble or map to with short reads [17]. Annotations of repetitive elements, such as TEs, in the reference genome is the first step when targeting this class of SVs and understanding their role in the formation of more complex SVs [80]. Long-read sequencing, such as Pacific Biosciences SMRT (PacBio) and Oxford Nanopore Technology (ONT) can help identify SVs and characterize breakpoints, especially in complex SVs [33].

Emerging methods for SV detection also include linked-reads, such as 10x Genomics, which provide long-range information across reads up to 100 kb or longer (e.g., [40]), or Strand-Seq, which preserves strand directionalities, but is mostly used in humans

(e.g., [81]). Chromosome conformation capture techniques like Hi-C provide long-range information at the chromosomal, and even inter-chromosomal, scale and are a powerful tool for characterizing complex SVs [46]. Compared to long reads, Hi-C data provide additional information about the potential effect of SVs on chromatin architecture, including enhancer-promoter contacts and consequent changes in gene expression [82], which is useful for linking genotype and phenotype. Optical mapping, based on visualization of restriction enzyme cut sites, or genetic mapping, based on linkage between genetic markers, are also valuable tools to validate large-scale SVs within or between chromosomes (e.g., [62,83]). Finally, comparison of *de novo* assemblies remains an important tool for SV detection, even within species, and can promote the creation of a pan-genome reference or a graph-based reference that includes major SVs from several individuals [6,84,85].

GLOSSARY

Amplified Fragment Length Polymorphism (AFLP): genomic marker obtained by amplification of a short fragment of DNA cut by restriction enzymes. Polymorphism is characterized by variable lengths.

Chromosomal inversion: a genomic structural variant in which a segment of DNA is reversed end-to-end relative to a reference sequence.

Copy number variant (CNV): a genomic structural variant in which a segment of DNA is represented in different numbers of copies. The segment can be absent (*deletion*) or present in two or more copies (*duplication[s]*) relative to a reference.

Expression Quantitative Trait Locus (eQTL): a genomic region that explains variation in mRNA transcript abundance.

Gene conversion: process by which one DNA sequence replaces a homologous sequence such that the sequences become identical after the conversion event.

Gene duplication: a genomic structural variant, example of CNV, in which a region of DNA that contains a gene is duplicated.

Haploblock (block of differentiation): region of reduced recombination, characterized by high LD, and often associated with high local differentiation between genetic groups.

Heterokaryotypes/homokaryotypes: individuals that are heterozygous/homozygous for a structural variant when it is considered as a single locus. The alleles are the different possible haplotypes (e.g., the inverted and non-inverted states for an inversion).

Insertion/deletion (indel): a genomic structural variant in which a segment of DNA varies in presence or absence relative to a reference. Indels include CNVs and non-reciprocal translocations.

Linkage Disequilibrium (LD): non-random association of alleles at different loci.

Microsatellites/minisatellites: a genomic structural variant, example of CNV, constituted by a tract of DNA motifs (1-10 bp for micro-, 10-60 bp for mini-) repeated 10 to 50 times. Also referred to as *tandem repeats* and *simple sequence repeats*.

Non-Allelic Homologous Recombination: a form of homologous recombination that occurs between two lengths of DNA that have high sequence similarity, but are not alternate alleles, such as TE copies.

Recombination suppression hypothesis: a model in which an inversion is indirectly favoured by natural selection because it suppresses recombination between sets of alleles, whereby alleles within a set are favoured in similar contexts and each set is favored in a different context.

Single Nucleotide Polymorphism (SNP): a single base-pair substitution.

Single Nucleotide Variant (SNV): genomic variant affecting a single base pair, including SNPs and single base-pair indels.

Structural Variant (SV): genomic variation between individuals affecting the presence, position, and/or direction of a nucleotide sequence (Figure 1).

Translocation: a genomic structural variant in which a segment of DNA is in a different position relative to a reference. Translocations can be either reciprocal or non-reciprocal (generating indels) and affect whole chromosome arms, such as in whole-arm reciprocal translocations***.*** The translocation of a segment of chromosome can result in a change in the total number of chromosomes, either by joining two chromosomes in one (*fusion*) or splitting a chromosome into two (*fission*). When fusions/fissions and translocations occur at the centromeres, they are called *Robertsonian****.***

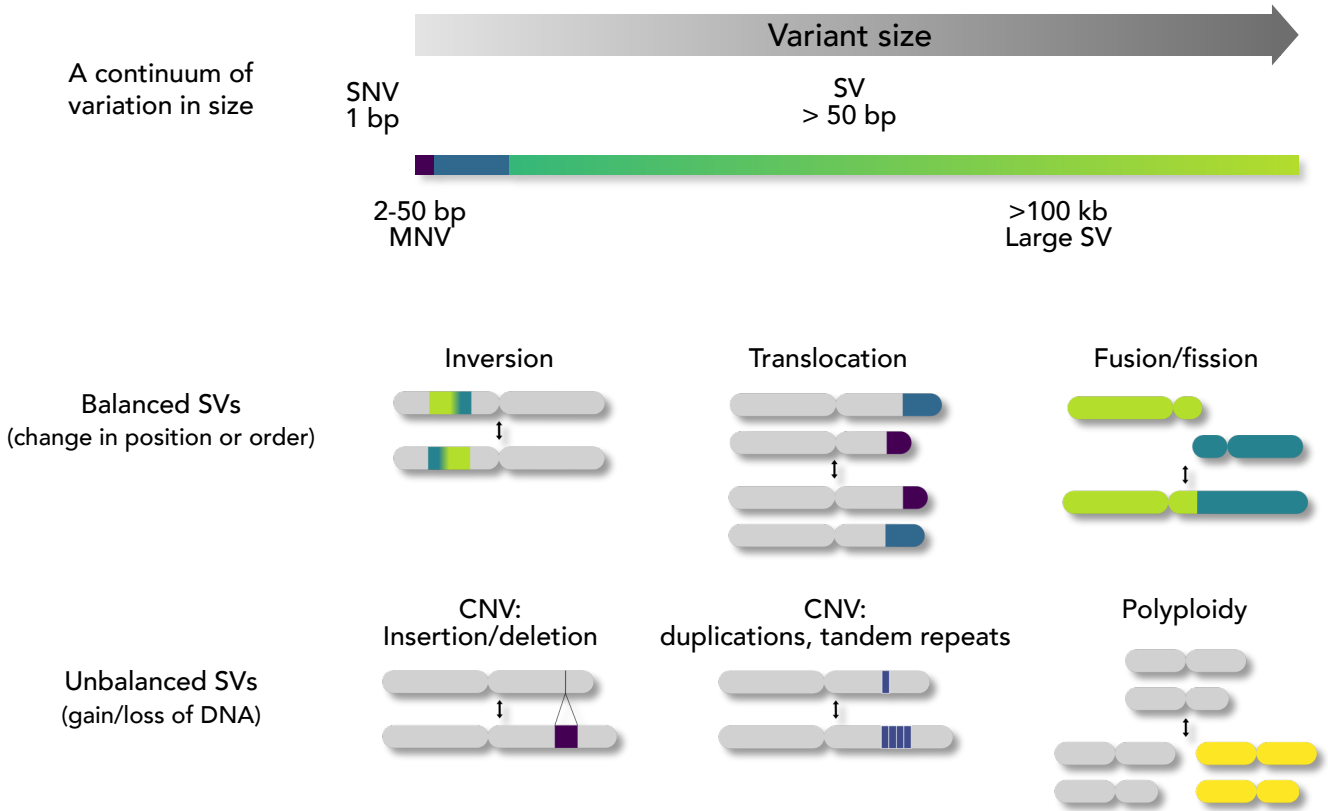
Transposable Element (TE or transposon): a segment of DNA that can change its position in the genome by either a cut-and-paste mechanism (DNA transposons) or a copy-and-paste mechanism (retrotransposons). TEs are a form of translocation, indel, and/or duplication.

HIGHLIGHTS

1. Structural genomic variants (SVs) take diverse forms and are ubiquitous drivers of ecological and evolutionary processes.
2. Most studies of SVs focus on the adaptive significance of gene duplications and large inversions. Future studies should catalogue SVs of all types and sizes and systematically test their evolutionary implications.
3. We propose a roadmap and definitions for the study of SVs in ecological and evolutionary genomics.
4. Best practices for SV detection are needed to facilitate comparisons across studies.
5. Integrating population genomic, theoretical, and experimental approaches to SVs will more comprehensively characterize genomic variation, uncover the adaptive and neutral processes shaping the evolutionary trajectory of SVs, and identify the mechanisms by which SVs impact adaptation and speciation.

OUTSTANDING QUESTIONS

- How can we develop appropriate bioinformatic tools to detect structural variants (SVs) of all sizes and genotype them in a large number of samples?
- What are the abundances, diversities, and distributions of SVs in natural populations and across taxonomic groups?
- How do SVs interact with sequence (e.g., SNP) variation and with each other? To what extent do different SVs predispose the offspring of carriers to more SVs?
- What are the roles of different types of SVs in evolutionary processes? For instance, which characteristics make some SVs particularly involved in adaptation and speciation? Conversely, how do neutral and adaptive processes determine the evolutionary trajectory of SVs?
- What is the relative influence of different types of SVs and sequence variation at different points along the speciation continuum and among systems with varying levels of gene flow?
- What are the proximate mechanisms (e.g., through linkage, effects on recombination, effects on 3D genome structure and gene expression, etc.) by which SVs influence evolution by natural and sexual selection?
- How can the unique properties of different types of SVs be harnessed for use as genetic markers to contribute to new understandings in population genomics and demography? What is the evolutionary rate of different SVs?
- How can SV markers be applied to agriculture, selective breeding programs, resource management, and conservation?


A continuum of variation in size
Variant size
SNV
1 bp
SV
> 50 bp
2-50 bp
MNV
>100 kb
Large SV
Balanced SVs
(change in position or order)
Inversion
Translocation
Fusion/fission
Unbalanced SVs
(gain/loss of DNA)
CNV:
Insertion/deletion
CNV:
duplications, tandem repeats
Polyploidy

Identification and characterization of SVs

Indirect methods
Patterns of differentiation and LD

Direct methods
Explicit mining of sequencing data

Objectives
Develop best practices to allow synthesis and comparison among studies
Assess sequencing/bioinformatic tools to select appropriate approaches depending on available resources
Account for population/species standing variation by enabling detection in large datasets

Understanding evolution and function of SVs

Population genomics
SV distribution & frequency
Differences among populations/species
Associations with environment and/or phenotype

Objectives
Acknowledge SVs as standing genetic variation
Describe evolutionary processes
Identify candidate SVs for adaptation & speciation

Evolutionary simulations
Simulate effects on genome evolution
Population genetic simulations of SVs distribution and evolution

Objectives
Predict evolutionary trajectory
Disentangle effects of neutral and selective processes on SVs

Experimental validation
Common garden and reciprocal transplant experiments
Experimental evolution
Differential gene expression
Gene knockdown/editing

Objectives
Validate candidate SVs for adaptation
Understand effects on phenotypes/recombination

Comparisons and meta-analyses among studies and across taxa

Comparative genomics - Phylogenomics
Synthesis of similarities and differences among studies and taxa

Objectives
Increase reproducibility and reliability of findings
Characterize SVs (e.g., number, position, size, breakpoints)
Assess frequency & distribution of SVs within and across species
Understand how SVs form, evolve, and persist

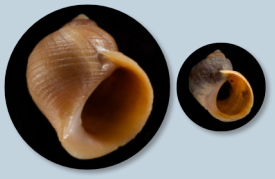
Ecological and evolutionary applications

New markers for genetic structure, environmental DNA, ancient DNA
Delineating evolutionarily significant units for conservation and management
Predicting population and species responses to global change
Agriculture and aquaculture breeding program design

GROUP

DIVERSIFICATION

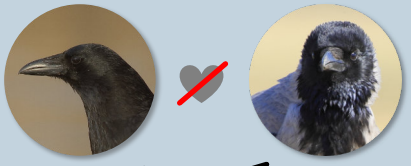
Local adaptation



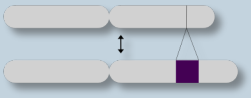
Littorina saxatilis
Wave/Crab ecotypes

17 inversions

Pre-mating isolation

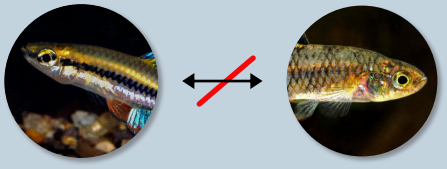


Corvus corone corone | *Corvus corone cornix*

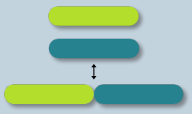


2.25-kb TE indel

Reproductive isolation



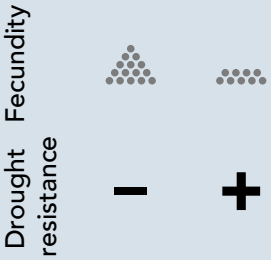
Lucania goodei | *Lucania parva*

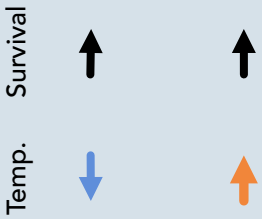


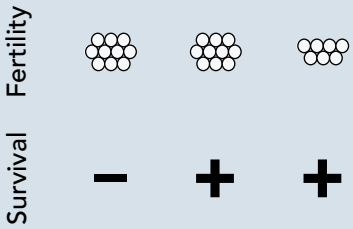
Robertsonian fusion

ORGANISM

FITNESS







PHENOTYPE

Flowering time

Mimulus guttatus

Plumage linked to thermal adaptation

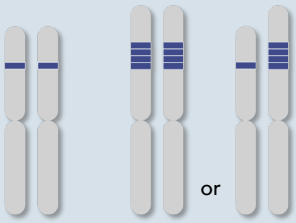
Uria aalge

Adult size & development time

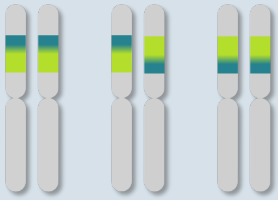
Coelopa frigida

GENOTYPE

CNV (x 300)



CNV (of a 60 kb region)

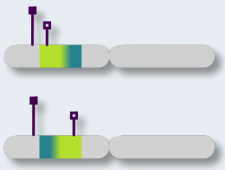


Inversion (25 Mb)

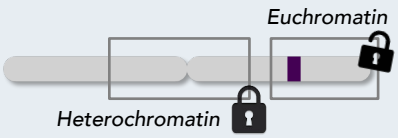
CELL

TRANSCRIPTOME

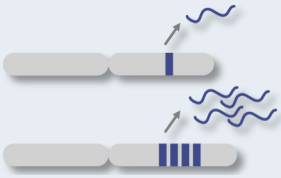
Gene-regulation decoupling



Position-effect variegation



Gene dosage

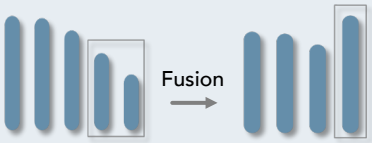


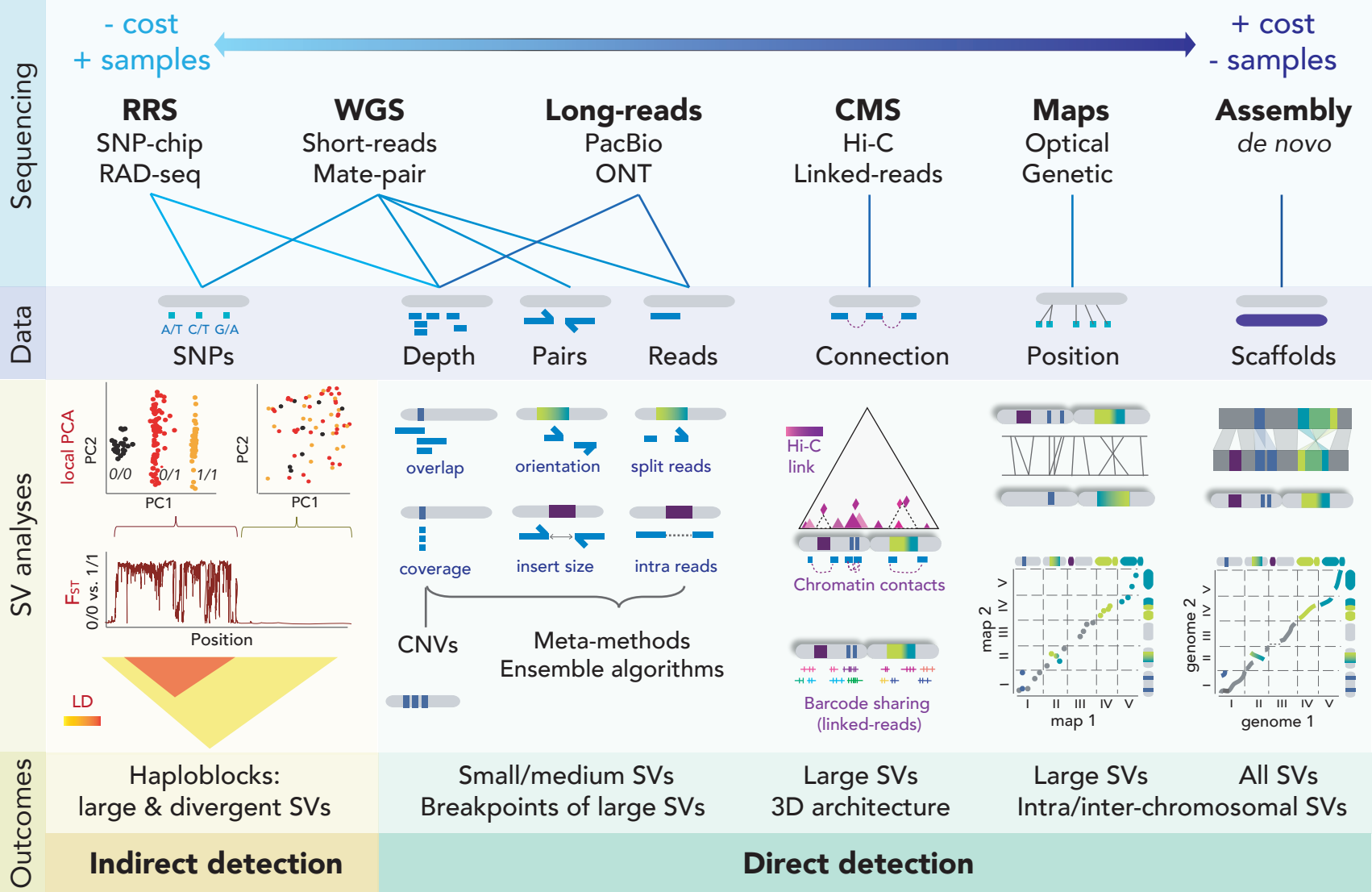
GENOME

Linear sequence

Recombination

Karyotype divergence




Sequencing
- cost
+ samples
+ cost
- samples
RRS
SNP-chip
RAD-seq
WGS
Short-reads
Mate-pair
Long-reads
PacBio
ONT
CMS
Hi-C
Linked-reads
Maps
Optical
Genetic
Assembly
de novo
Data
A/T C/T G/A
SNPs
Depth
Pairs
Reads
Connection
Position
Scaffolds
SV analyses
local PCA
PC2
PC1
0/0
0/1
1/1
PC2
PC1
FST
0/0 vs. 1/1
Position
LD
overlap
orientation
split reads
coverage
insert size
intra reads
CNVs
Meta-methods
Ensemble algorithms
Hi-C
link
Chromatin contacts
Barcode sharing
(linked-reads)
map 2
map 1
genome 2
genome 1
Outcomes
Haploblocks:
large & divergent SVs
Small/medium SVs
Breakpoints of large SVs
Large SVs
3D architecture
Large SVs
Intra/inter-chromosomal SVs
All SVs
Indirect detection
Direct detection