The potential for next-generation sequencing to characterise the genetic variation underlying non-syndromic cleft lip and palate phenotypes

A Collins1*, L Arias2, R Pengelly1, J Martínez2, I Briceño2, S Ennis1

Abstract

Introduction

Next-generation sequencing is revolutionising the study of genetic variation and its role in disease. Individual DNA samples can now be sequenced cost-effectively enabling analysis of the complete spectrum of genetic variation. This technology has the potential to contribute significantly to the understanding of non-syndromic cleft lip and/or palate. This condition occurs with relatively high frequency and only a proportion of the underlying genetic causal factors have been identified. Many of the genes implicated have been found through genome-wide association studies but further progress is limited because these approaches consider only common genetic variants and neglect rarer variants. Because many of the causal genetic variants remain unknown, the role of gene-environment and gene-gene interaction is difficult to characterise. The identification of novel, low frequency, variants will provide new insights into the biological mechanisms and pathways involved in the condition. Sequence-based analysis will also be invaluable for fine mapping causal variants in the larger regions already identified by linkage and association studies for which positive identification of causal genetic variants has proven difficult. This review considers the available evidence for the genes involved and current understanding of how genetic variation interacts with environmental factors known to influence risk. Only by characterising the underlying genetic factors will the effort to understand gene-environment interaction and underlying functional processes be successful.

Conclusion

Success with next-generation sequencing will lead to improvements in prediction, prevention, and treatment for cleft lip and palate patients.

Introduction

Next-generation sequencing (NGS) is revolutionising genomics and this trend of increasingly significant impact is likely to continue. Most importantly, NGS provides a route to characterising and understanding the role of genetic variation in disease. This is important because evidence from genome-wide association studies (GWAS) suggests that much of the heritability underlying complex disease phenotypes is not explained by common deleterious genetic variants with small effect sizes. NGS enables new analytical strategies, which are not achievable through genome-wide association studies. These include: the identification of the complete complement of DNA variants in samples (across the allele frequency spectrum); tests on the burden of rare variation (do specific genes contain many rare variants which collectively impair gene function?); the identification of de novo mutations and the genes underlying rare Mendelian forms of disease; fine mapping of causal variants within broader regions identified by linkage and/or association and the characterisation of important structural variation, such as differences in copy number, which may contribute to disease.

Orofacial cleft lip and/or palate (CLP) represents a complex phenotype for which NGS offers the potential to increase understanding. CLP phenotypes are among the most frequent birth defects with rates of between 1/500 and 1/2500 births. The frequency of CLP phenotypes is related to population ancestry, geographical location, maternal age, prenatal exposures and socioeconomic status. The frequency of orofacial clefting (OC) is higher in Latin American and Asian countries. CLP phenotypes are classified into syndromic and non-syndromic forms. The former includes many conditions which have simple Mendelian modes of inheritance in families for which a number of causal genes have already been identified through, for example, linkage mapping. However, ~70% of CLP cases occur as isolated phenotypes without any additional cognitive or craniofacial structural abnormalities. These are usually described as isolated non-syndromic cleft lip and/or palate (NSCLP). Understanding the factors underlying NSCLP phenotypes is important to improve prevention, treatment and prognosis of the condition. However, the genetic dissection of NSCLP phenotypes is challenging and progress towards understanding the underlying genetic and environmental factors, and how they are inter-related, has, until recently, been relatively slow, despite decades of research. This review considers the impact of recent work and, in particular, prospects for progress through...
the application of NGS to further characterise underlying genetic variation and its role in NSCLP.

**Genetic factors and NSCLP**

NSCLP is a genetically complex disorder which results from interactions between multiple genetic and environmental risk factors. The disorder has a significant genetic basis and it is known that first degree relatives of affected individuals have a 30–40 fold increased risk compared to the background population. The degree of phenotype concordance for monozygotic (MZ) twins is 40–60% compared to 5% for dizygotic twins. Murray and Grosen et al. found heritability estimates exceeding 90% for CLP phenotypes. Genetic studies including linkage analysis, genome-wide association, and GWAS-based meta-analysis, have yielded reproducible evidence for several genes and gene regions. Results from Ludwig et al. identified four genes and gene regions (IRF6, 8q24, 17q22 and 10q25; Table 1) for which the total population attributable risk is ~55% suggesting that, unusually for a complex trait, a substantial proportion of the variation in NSCLP might be explained by these loci. However, many uncertainties remain. Poor concordance between regions identified by linkage with those found by association mapping (Table 1) must reflect in part the different targets of the techniques. Association mapping is good for detecting common variants contributing small effect sizes in population samples whereas linkage mapping is more powerful where there is allelic heterogeneity, for example where multiple (rare) variants in a particular gene contribute to disease. But, although many of these signals have been replicated in independent samples, several of the linkage regions are broad and the underlying causal gene(s) are poorly established. Incomplete knowledge about gene function presents difficulties for selecting the most likely

<table>
<thead>
<tr>
<th>Nearest or causal gene</th>
<th>Region</th>
<th>Protein function</th>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAX7</td>
<td>1p36</td>
<td>Transcription factor: neural crest development in mouse</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>ARHGAP29</td>
<td>1p22</td>
<td>Regulation of binding proteins involved in craniofacial development</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>IRF6</td>
<td>1q32</td>
<td>Involved in formation of connective tissue</td>
<td>Association/ linkage</td>
<td>12–14</td>
</tr>
<tr>
<td>THADA</td>
<td>2p21</td>
<td>Possible regulatory functions</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>TGFA</td>
<td>2p13</td>
<td>Involved in signalling pathway for cell proliferation, differentiation and development</td>
<td>Linkage</td>
<td>12</td>
</tr>
<tr>
<td>EPHA3</td>
<td>3p11</td>
<td>Regulation of cell shape and cell:cell contacts</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>-</td>
<td>8q21</td>
<td>Intergenic region</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>-</td>
<td>8q24</td>
<td>Gene desert: may contain regulatory elements for craniofacial development</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>FOXE1</td>
<td>9q21</td>
<td>Transcription factor regulating diverse developmental processes</td>
<td>Linkage</td>
<td>12, 13</td>
</tr>
<tr>
<td>VAX1</td>
<td>10q25</td>
<td>Transcriptional regulator</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>SPRY2</td>
<td>13q31</td>
<td>Signal is inter-genic, nearest gene is regulator of multiple receptor tyrosine kinases</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>PAX9,TGFβ3,BMP4</td>
<td>14q</td>
<td>BMP4: bone morphogenetic protein involved in bone/cartilage development</td>
<td>Linkage</td>
<td>12, 13</td>
</tr>
<tr>
<td>TPM1</td>
<td>15q22</td>
<td>Signal is inter-genic, in regulatory region, gene encodes actin-binding protein</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>FOXC2, CRISPLD2</td>
<td>16q24</td>
<td>FOXC2: transcription factor, possible role in development of mesenchymal tissues</td>
<td>Linkage</td>
<td>12, 13</td>
</tr>
<tr>
<td>NOG</td>
<td>17q22</td>
<td>Essential for cartilage morphogenesis and joint formation</td>
<td>Association</td>
<td>14</td>
</tr>
<tr>
<td>MAFB</td>
<td>20q12</td>
<td>Transcription factor involved in development of keratinocytes</td>
<td>Association</td>
<td>14</td>
</tr>
</tbody>
</table>
candidate gene(s) in these regions. Several gene regions identified by linkage in earlier studies have not replicated subsequently in independent samples. Successful replication is difficult to achieve and it is perhaps too early to dismiss some of the more uncertain signals. Association mapping frequently reveals variants in inter-geric regions which are suggested to have regulatory functions influencing gene(s) nearby. Such a mechanism has only been firmly established for a small number of regions and identifying the precise causal variant(s) is made more difficult because of extensive linkage disequilibrium. It is particularly difficult to understand the precise functional roles of these apparent regulatory variants. Frequently, the nearest gene and/or gene with the most plausible NSCLP-related function is highlighted (Table 1). Other issues which have not been resolved by linkage and association studies include causes of apparent differences in the underlying genetic basis of NSCLP between populations of different ethnicity. One region which has been extensively studied is 8q24 for which Murray et al. found much stronger evidence in European-derived samples, compared to Asians. In this case, the difference was attributed to reduced haplotype diversity in the Asian sample reducing power, rather than a distinct genetic effect. It is far from clear that such a mechanism accounts for ethnic genetic differences in other candidate gene regions.

Although understanding of the genetic basis of NSCLP phenotypes has advanced considerably in recent years, many unanswered questions remain, for which NGS may offer a route to progress. NGS has advanced considerably in recent years. NGS has the potential to help determine actual causal variants rather than assignments to a broader region. The sequencing of many NSCLP genomes will be essential to establish models which consider the roles of regulatory sequences and the genes involved.

Environmental factors and gene-environment interaction

Although high MZ concordance is consistent with substantial genetic influence, the incomplete concordance suggests non-genetic influences on NSCLP phenotypes. Environmental effects might generate incomplete penetrance through random developmental events or a non-homogeneous in utero environment. Grosen et al. found MZ twin discordance might reflect genetic, cytogenetic or epigenetic anomalies in the affected twin that are not found in the unaffected twin. Post-zygotic genomic alterations resulting from mitotic recombination have been considered but have been shown by Kimani et al. to not be a common cause of MZ twin discordance in CLP. Their analysis did not exclude rare or balanced genomic alterations, tissue-specific events and small aberrations beyond the resolution of their methods (~1Mb). Sequence-level resolution achieved by NGS might be informative given appropriately designed studies.

Establishing relationships between genetic and environmental factors has proven extremely challenging so far. Skare et al. conducted a large study aimed at detecting interactions between 334 candidate genes and maternal first trimester exposure to smoking, alcohol, coffee, folic acid supplements, dietary folate and vitamin A. This study contrasted 425 case-parent triads with 562 control-parent triads. Very little evidence for gene-environment interaction was found in these data. They noted that ‘it is remarkable that OC, a phenotype of supposedly very high heritability, remains so hard to decipher’. The authors consider that larger sample sizes and, therefore, greater power to establish effects are required. Butali et al. examined interactions between the MTHFR gene C677T variant and folic acid in OC aetiology. They contrasted 1149 isolated cases and 1161 controls and considered maternal peri-conceptional exposure to smoking, alcohol and folic acid. Although folic acid and smoking were found to influence OC outcomes, no significant interaction was demonstrated with the C677T variant. Beaty et al. found some evidence for gene-environment interaction using available data on maternal smoking during pregnancy in European case-parent trios. The genes involved were GRID2 and ELAVL2. However, neither gene showed evidence of association with NSCLP in the absence of the smoking interaction effect.

Efforts have been made to understand the underlying molecular mechanisms behind NSCLP and their relationship to genetic and environmental factors. Studies contrasting the transcriptome of dental pulp stem cells from NSCLP patients with controls suggest that there are alterations in gene networks (differentially expressed genes) functionally relevant to orofacial development, such as collagen metabolism and extracellular matrix remodelling. Because NSCLP is considered to arise through anomalies in cellular migration, proliferation, trans-differentiation and apoptosis. Kobayahi et al. considered possible overlap between NSCLP and cancer gene pathways. They demonstrated that NSCLP patient-derived stem cells show dysregulation in gene networks controlling cellular defences against DNA damage. The authors speculate that alterations in a small number of upstream genetic or epigenetic regulators, combined with deleterious genetic variants could disrupt the modulating activity of transcription factors such as E2F1. Hence genetic and epigenetic variation underlying regulatory anomalies, combined with environmental factors, may be driving NSCLP. Continuing progress with
this functional work is hampered, in part, by incomplete understanding of underlying genetic factors. Specifically, it is not clear that the small number of NSCLP variants identified thus far can account for the dys-regulation of cellular functions and pathways identified and none of the differentially expressed genes identified correspond to the known GWAS variants. To investigate how regulatory anomalies underlie the development of NSCLP, it is necessary to characterise the complete spectrum of variation in genome sequences, including all regulatory variants in non-coding regions.

NGS: some practical considerations

Although NGS has enjoyed dramatic success through the identification of genes underlying Mendelian disorders, and also de novo disease mutations, complex phenotypes such as NSCLP have proven much more difficult to elucidate. Aside from numerous data quality control, technical and data management issues, a particular difficulty arises from the many, often apparently deleterious DNA variants, identified in each DNA sample. Faced with this complexity, various methods to ‘filter’ variants lists are undertaken to try to exclude ‘neutral’ variation. This procedure involves removal of ‘common’ variants (those represented in high frequency in databases of sequence variants from individuals lacking recognised disease) and removal of implausible disease candidates in genes which are highly mutable. Such genes include those with sensory or immune functions for which high allele diversity is adaptive. Frequently, it is cost-effective to sequence only the protein coding exons of genes (the ‘exome’), representing only 1% of the genome. From an exome sequence, non-synonymous variants (those that change an amino acid in the protein) can be selected for further study and other variants excluded. A disadvantage of using only the exome and extensive filtering is that variants in non-coding regions, which may have regulatory functions, along with much of the structural variants, are excluded. Even a highly filtered list of non-synonymous variants may contain many potentially deleterious variants, which do not in fact influence the phenotype. Various predictive metrics such as SIFT and PolyPhen2 have been developed which help discriminate potentially deleterious variants from those that are neutral. SIFT predicts whether an amino acid substitution affects protein function based on conservation of amino acid residues across species. PolyPhen2 considers impacts of an amino acid substitution on the structure and function of a protein. Low scores (~0) for SIFT and high scores (~1) for PolyPhen2 suggest that the variant may be deleterious and contribute to disease. Figure 1 presents SIFT and PolyPhen2 scores for non-synonymous variants in the IRF6 gene, which contains variants involved in both syndromic forms of CLP (Van der Woude and popliteal pterygium syndrome) and NSCLP (Table 1). Scores for variants in this gene from the Exome Variant Server (EVS: a database of 6400 exome sequences) and known disease causal variants from the Human Gene Mutation Database (HGMD) are shown. The score for an exome-sequenced patient from Colombia with typical popliteal pterygium syndrome, who has the rs121434226 single-nucleotide polymorphism in IRF6, is also given. Although there is a degree of separation between known neutral and known causal variations (and the Colombian patient score is clearly deleterious by both measures), there is also overlap. These functional predictive methods can be useful for ranking variants worthy of further investigation but are not fully discriminatory, particularly for complex phenotypes where individual variants have reduced penetrance.

Figure 1: Polyphen2 and SIFT scores of variants in the IRF6 gene. PolyPhen-2 versus SIFT scores for non-synonymous variants in the IRF6 gene. Presumed neutral variants from the Exome Variant Server (EVS, n = 12), variants reported to cause CLP from the Human Gene Mutation Database (HGMD, n = 80) and the aetiological SNP from an exome sequenced in our Colombian patient (Col) are shown. Variants known to cause CLP are clearly clustered in the bottom right of the plot, representing a predicted deleterious nature by both metrics.
Discussion

The identification of genetic factors underlying NSCLP has proven extremely challenging although recent progress with GWAS, and subsequent meta-analyses, have firmly implicated a number of genes and variants in NSCLP phenotypes. Although multiple loci identified through GWAS appear capable of explaining a relatively high proportion of the heritability low concordance between linkage and association studies strongly suggests that rarer variations, which can be detected by NGS, will provide additional causal insights. NGS sequencing studies will be invaluable for fine mapping causal variants in linkage and GWAS-identified genes and in pursuing additional, rarer; variations in related genes and pathways, along with novel genes. Only by developing a greater understanding of the underlying genetic basis of NSCLP will efforts to understand gene-environment interaction and functional processes underlying NSCLP be successful.

NGS also has the potential to contribute to understanding of the roles of different genetic factors amongst different ethnic groups and how these interact with diverse environmental influences. NGS is also capable of delimiting distinct disease sub-types within the NSCLP ‘umbrella’ which is important for refining diagnosis and tailoring treatment. The development of integrated models which consider gene-gene and gene-environment interaction and how these influence the function of key pathways will underpin more complete understanding. Although exome sequencing is valuable, whole genome sequencing of many individuals from different populations, comprehensive phenotyping, and careful consideration of environmental factors, may be required for establishing regulatory roles of some variants. However, NGS presents considerable challenges for data analysis and interpretation. Much effort is now focussed on addressing these difficulties and, as many more genomes are sequenced, further success in understanding the role of genes in NSCLP phenotypes is expected.

Conclusion

Although important recent advances have revealed some of the genetic variants underlying NSCLP, NGS has the potential to identify novel genetic factors. Only given more comprehensive understanding of genetic variation underlying NSCLP can interactions between genes, and between genes and environmental variables, be firmly identified. Success with NGS will lead to improvements in prediction, prevention and treatment for cleft lip and palate patients.

Acknowledgements

The authors gratefully acknowledge funding from the Newlife Foundation for Disabled Children.

References


Critical review

Competing interests: none declared. Conflict of interests: none declared. All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript. All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.